SISSA

Scuola Internazionale Superiore di Studi Avanzati

Trieste, Italy

Neuroscience Area

PhD course in Cognitive Neuroscience

Texture Representations in Deep Neural Models and Humans

Supervisor:
Eugenio Piasini

Candidate:
Ludovica de Paolis

ACADEMIC YEAR 2025/2026



Acknowledgments

I would like to thank Eugenio for being my supervisor and a mentor. I started my PhD having no idea on how to even turn on a laptop, but you encouraged me to learn and supported me in facing the many obstacles I encountered. I will be forever grateful for the trust you put in me. Thanks also to Fabio and Alessio, for the scientific guidance and the collaborations, especially during my first project. Thanks to Alessandro and Marco for "adopting" me, for the genuine interest and enthusiasm about my research. Marco, thank you for hosting me in Barcelona. Alessandro, thank you for taking me to the mountains and the politics discussions. I would also like to thank the commission for their availability and interest in participating to my defense: Diego Doimo, Francesca Franzon, Sebastian Goldt, Daniel Kaiser, and Alex Rodriguez. Thanks to Beatriz for the valuable human and academic advice.

This thesis would not be complete without the precious and unconditional support of the people I encountered in the last four years. Thanks to the neurostocastici, Andrea, Gabri, Kostas, Marcin, Marlen. You were there from day zero, and you are still here today. Thank you Gabri for the philosophical walks in the sunset and the dreams; Kostas for the "pedimou" and for patting my shoulder vigorously; Marcin for the many dances together; Marlen for the dinners, the grappini, the cats, and the sewing. Thanks to Anna, for sharing our first office, the beauties and the struggles of being a woman in academia. Thanks to Emma (Juan) for being such an important part of my life in Trieste, for always being there when things got tough and dramatic. Thanks to Ila and Smema, for all the sleepovers and for being present even when our job made us apart. Thanks to Giulio for three years of support, affection and freedom by my side. Thanks to Aldino, for being the best flatmate I could ever ask for. Thanks to Elisa for all the ice-cream, music, climbs, and dinners together. Thanks to Deb, for your friendship and support, for the activism, and for your questionable humor. Thanks to Younes, for all the food and for the strolls in weird places. Thanks to Mini, Moha, and Nour, for your kindness, compassion, and dedication. Thanks to all the people who supported me while I was representative in the Senate: Alice, Alessandro, Andrea Ros., Emma D., Giammi, Giorgos, Maria A., Matteo, Max, Nicoletta, Salvo, Youssef. Thank you for taking a stance and for expressing your opinion, especially in the hardest moments. Thanks to my climbing mates, in particular Ece, Fedo, Gianni, Marco, Matéo. To Gianni, thanks for all the expeditions to Sampeter; to Ece and Matéo, thanks for all the beautiful moments in Barcelona. Thanks to Mauro for the memes and the nonsense; to Micfossati for the hikes and the tektonik; to Carlo, Daniele, Lindon, Maxim, Paolo, and Tadeo for being amazing lab and office mates; to the COLT people for welcoming me. Thanks to the SISSA mensa staff, you made me smile even in the worst days. Thanks to Elisa and Lucio for taking care of me. Thanks to everyone, scientist or not, who supported the cause of Palestine. Finally, to my family: to my brother Egidio, and to my parents, Elisabetta and Nuccio. You always believed in what I was doing, supported every choice and every journey to the other side of the world. You embraced any decision that I took. You traveled and called me for hours, just to be close to me when I struggled to keep everything together. I will be forever grateful for all of this and much more. Thank you.

Abstract

Visual perception is often studied through the lens of object recognition. However, a large fraction of visual experience concerns image properties that are not naturally described in terms of discrete objects or categories. Visual textures provide a useful method to investigate this complementary aspect of vision, as they can be described through image statistics while spanning a wide range of perceptual and semantic complexity. In this thesis, I investigate how textures are represented in deep neural networks and humans, combining texture synthesis, representation analysis, computational modeling, and psychophysics.

I first explored whether the naturalistic texture representations introduced by Gatys et al. (2015) could be combined with the latent space of a Variational Autoencoder to construct a controllable generative model of naturalistic textures. Although the complete model did not achieve the intended performance, the experiments isolated the texture representation proposed by Gatys as the most robust component of our framework and motivated a direct investigation of its representational properties.

I then studied texture representations coded as Gram matrices across a diverse set of convolutional neural networks trained for object recognition. Texture information increased with network depth, but the resulting representations recovered only part of the categorical structure of natural textures defined by human judgment. Crucially, their quality was not correlated with measures of neural or behavioral alignment in object recognition. This dissociation indicates that being a good model of the ventral visual stream in an object recognition setting does not necessarily imply representing textures in a human-like way.

Finally, I extended this analysis beyond convolutional architectures by comparing one CNN with three Vision Transformers across textures of progressively increasing complexity and with human performance in a psychophysical discrimination task. Vision Transformers formed substantially more consistent representations across texture types than the CNN, and these representational relationships were significantly better aligned with human perception.

Taken together, these results show that texture perception provides a complementary benchmark for computational models of vision and exposes limitations within the dominant object recognition paradigm. More broadly, I suggest that understanding biological visual representations requires models and evaluation criteria that account not only for object identity and semantic content, but also for the statistical and contextual structure through which visual appearance is organized.

Contents

Abstract **5**

List of Publications **9**

List of Figures **13**

List of Tables **16**

1 Introduction **1**
1.1 The ventral visual stream . . . 2
1.2 Computational models of the ventral visual stream . . . 4
1.3 Artificial vision . . . 7
1.4 Visual textures . . . 12
1.5 Models for visual textures synthesis . . . 15
1.5.1 Victor and Conte . . . 16
1.5.2 Portilla and Simoncelli . . . 19
1.5.3 Gatys . . . 21

2 A Variational Autoencoder for texture synthesis **31**
2.1 Abstract . . . 31
2.2 Introduction . . . 32
2.3 Methods . . . 34
2.3.1 Model architecture . . . 34
2.3.2 Loss . . . 34
2.4 Experiments . . . 35
2.4.1 Evaluation of the Gatys algorithm . . . 35
2.4.2 Benchmarking datasets . . . 36
2.4.3 Large unlabeled datasets . . . 37
2.4.4 Texture dataset . . . 38
2.5 Discussion and conclusions . . . 40
2.6 Appendix . . . 41
2.6.1 Model architecture . . . 41
2.6.2 Experimental configurations . . . 44
2.6.3 Supplementary figures . . . 46

3 Perceptual misalignment of texture representations in convolutional neural networks **53**
3.1 Abstract . . . 53
3.2 Introduction . . . 53
3.3 Methods . . . 55
3.3.1 The “Gram representation” for textures . . . 55
3.3.2 Texture synthesis . . . 55
3.3.3 CNN Layer selection . . . 56
3.3.4 Stimulus dataset . . . 56
3.3.5 Clustering . . . 57
3.3.6 Measuring clustering quality . . . 58
3.3.7 Brain-Score correlation . . . 59

3.4 Results . . . 59
3.4.1 Representational Similarity Analysis . . . 59
3.4.2 Clustering analysis . . . 61
3.4.3 Texture synthesis . . . 62
3.5 Discussion . . . 62
3.6 Conclusions . . . 64
3.7 Appendix . . . 65
3.7.1 Loss normalization . . . 65
3.7.2 Supplementary Tables . . . 66
3.7.3 Supplementary figures . . . 70

4 Texture Representations in Deep Vision Models: Comparing CNNs, Vision Transformers, and Human Perception 81
4.1 Abstract . . . 81
4.2 Introduction . . . 81
4.3 Methods . . . 83
4.3.1 Stimulus datasets . . . 83
4.3.2 Models . . . 84
4.3.3 Human psychophysics . . . 85
4.3.4 Information Imbalance . . . 86
4.4 Results . . . 86
4.5 Discussion . . . 90
4.6 Limitations . . . 91
4.7 Appendix . . . 95
4.7.1 Supplementary analysis . . . 95
4.7.2 Supplementary figures . . . 98
4.7.3 Stimulus dataset generation . . . 101
4.7.4 Models details . . . 103
4.7.5 Human behavioral experiment . . . 103

5 Conclusions 109

List of Publications

- Ludovica de Paolis, Fabio Anselmi, Alessio Ansuini, Eugenio Piasini *Perceptual misalignment of texture representations in convolutional neural networks* [Accepted in Journal of Vision]
- Ludovica de Paolis, Marco Baroni, Alessandro Laio, Eugenio Piasini *Texture Representations in Deep Vision Models: Comparing CNNs, Vision Transformers, and Human Perception.* [arXiv preprint]

List of Figures

1.1 **Architecture of the Vision Transformer**, adapted from Dosovitskiy et al. (2021), Figure 1. . . . 11
1.2 **The 16 possible binary 2×2 configurations.** . . . 17
1.3 **10-dimensional parametrized space of textures** defined by the algorithm by Victor and Conte (2012). The horizontal axis ranks statistics by increasing order, starting from first-order. Adapted from Hermundstad et al. (2014), Figure 2A. . . . 18
1.4 **Comparison between similar textures** synthesized in the original paper by Victor and Conte (2012) (a), and with Metex (b). . . . 18
1.5 **Diagram of the recursive texture synthesis algorithm.** Adapted from Portilla and Simoncelli (2000a). . . . 20
1.6 **Comparison between textures synthesized** by Portilla (Figures 1.6a, 1.6c), and with the algorithm implementation by Portilla and Simoncelli (2000b) (Figures 1.6b, 1.6d), showed in the bottom row. The top row shows two input images used in the original work by Portilla and Simoncelli, and two input images from the Describable Texture Dataset (DTD) *pitted_0008.jpg*, *pitted_0025.jpg* (Cimpoi et al., 2014). Adapted from Portilla and Simoncelli (2000a) Figures 14, 15. . . . 20
1.7 **The Gatys model.** Adapted from Figure 1 in Gatys et al. (2015). . . . 21
1.8 **Representation of a feature map** of dimensionality $F^{H \times W \times M}$, where $M = 4$, and its Gram matrix of dimensionality $G^{M \times M}$. The entries i, j of the Gram matrix are the dot products between all the possible combinations of the vectorized channels of the feature map. . . . 22
1.9 **Textures generated artificially with the Gatys model.** Original implementation vs. my implementation. Figure 1.9a shows the original input image, a picture of pebbles; figure 1.9b shows the synthesized texture by Gatys et al. with the original implementation and the optimization algorithm LFBGS; figure 1.9c shows the synthesized texture with my implementation with Adam optimizer at 130000 steps; figure 1.9d was synthesized with my implementation and LFBGS at 6000 steps. Adapted from Gatys et al. (2015) and de Paolis et al. (2026). . . . 23

2.1 **Replication of the Gatys algorithm** with two original images (first column) and corresponding reconstructions (second column). . . . 36
2.2 **MNIST and CIFAR-10 training experiment.** Random samples of original images (first column) and reconstructions (second column). This figure is for illustrative purposes to examine the quality of the output with respect to the quality of the input. . . . 37
2.3 **Custom dataset and NSD training experiment.** Random samples of original images (first column) and reconstructions (second column). This figure is for illustrative purposes to examine the quality of the output with respect to the quality of the input. . . . 39
2.4 **DTD training experiments.** Original images (first column) and corresponding reconstructions (second column). In both the AE and the VAE versions of the model I trained for 100 epochs and kept the parameter configurations displayed in Table 2.1. . . . 40

2.5 **Diagram of the model's architecture.** The backbone is a VAE (top), equipped with two VGG-19 instances (middle) to extract the Gram matrices (bottom). 42

2.6 **Losses landscapes of our model trained with NSD** after 2 training days. 46

2.7 **Losses landscapes of VAE trained with DTD** at 10.000 epochs. . . . 47

3.1 **Representational Dissimilarity Matrices** obtained from the five layers analyzed for VGG-19. The size of each matrix is 5640×5640, and each entry corresponds to the Euclidean distance – relative to each matrix, as in the colorbars – between the Gram matrix representation of one pair of images in DTD. The entries are sorted such that images belonging to each of the 47 ground-truth categories are placed next to each other. 60

3.2 **MI values across CNN layers** identified by index (1-5). Each line corresponds to one of the 12 models as color-coded in the legend. 61

3.3 **Plots of the correlations with Pearson's *p*.** Each subplot represents the correlation of Brain-Score benchmarks *average vision, neural vision, behavior vision, V1, V2, V4, IT* respectively vs. the layer with the highest MI for each of the 12 CNNs. Each color identifies a model as reported in the legend. 62

4.1 **Example stimuli arranged along a putative spectrum of texture complexity.** In the Noise sample, each pixel is sampled independently at random as black or white with equal probability. The stimuli labeled as V&C, P&S, and G are obtained by the DTD image classified as "braided" (marked as "source") and by application of the respective algorithm. The Object image was chosen by hand for illustrative purposes from the Object dataset. 85

4.2 **Graphs representing the symmetrized mean II across models.** For visualization purposes, for each subset we selected the best II between the two most mutually predictable layers for all models. For ease of illustration, we used $1 - II$: strong links and high values show high mutual predictability; thin links and low values represent low predictability. An extended figure is available in Appendix 4.7.2, Figure 4.7. 87

4.3 **Graphs representing the symmetrized mean II across datasets**, excluding Object. For representational purposes, for each model we selected the II between the best mutually predictive layers of all the subset combinations. For illustrative purposes, we inverted the II scale to $1 - II$: strong links correspond to high mutual predictability; thin links represent low predictability; slightly thick links represent equal mutual predictability. An extended figure is available in Appendix 4.7.2, Figure 4.8. 88

4.4 **Human pairwise accuracy and its correlations with models performance.** For illustrative purposes, in panel (a) we represent $1 -$ accuracy, so that thicker links mean stronger perceptual overlap. Panel (b) shows the four correlation plots among humans and the four models. We used Pearson's r and we report *r, p* inside the boxes. The legend maps a color for the II score used in the correlation for each pair of image types. 89

4.5 **Heatmaps for each model (rows) by each dataset (columns).** For each matrix, the entries are the symmetrized mean II values calculated between all the possible combinations of layers. Each matrix is equipped with its relative color-bar of II values in logarithmic scale. 97
4.6 **Linear probe experiments. (a) Classification**: accuracy scores for a triplet of probes trained on early, middle, and deep extracted features for each dataset. **(b) Transfer learning**: accuracy scores for a triplet of probes trained on early, middle, and deep layers' extracted features of DTD and tested on the other datasets. 98
4.7 **Extended plot for Section 4.2, Figure 4.2**: For each dataset (columns) we compute II for each pair of models. In particular, as shown in the rows, each model serves as a *reference*, and the other three serve as *target*, as shown in the columns and in the top legend. We quantify the depth of each feature's layer as the relative percentage, and compared symmetrized II scores between reference layers and target layers along the same relative depth. 99
4.8 **Extended plot for Section 4.3, Figure 4.3**: For each model (rows) we compute II between all pairs combination between a *reference dataset* (columns) and four remaining *target datasets* (legend). We quantify the depth of each feature's layer as the relative percentage, and compare symmetrized II scores between reference layers and target layers along the same relative depth. 100
4.9 **Extended plot for Section 4.5, Figure 4.4a**: This plot represents the relationships among texture subsets for each pair of *odd image* (see X axis) and and *reference image* (legend) as human accuracy scores (see Y axis) of all participants collected in the behavioral task. Each box summarizes the distribution of mean accuracies for one combination of odd source and reference source. The scatter shows the individual data points, the box shows the interquantile range, the central line shows the median, and the errorbars show the non-outlier range. We observe that accuracy is highly odd source-dependent. Trials with V&C as the source were solved with almost top accuracy for all odd-source conditions, indicating that V&C images form a perceptually coherent and distinct subset. In contrast, when V&C was the odd source, performance was moderate: accuracy in distinguishing P&S, G and DTD gets lower, suggesting a substantial perceptual overlap. This asymmetry suggests that odd-one-out performance is not determined by a symmetric perceptual distance between datasets, but also by the internal variability of the reference source. 101
4.10 **Visual stimulus screen in the experimental design.** The odd in this case is the bottom stimulus (↓). 104
4.11 **Illustration of the experimental flow**, with time in ms. Each simulation is paced at 200ms, proceeded by a 400ms red fixation cross. Response time is self paced. Breaks between trials are provided. 104

List of Tables

1.1 **Architectures of VGG-16 and VGG-19.** Adapted from Simonyan and Zisserman (2015), Table 1. 9

2.1 **Architecture of the Variational Autoencoder backbone.** 41
2.2 **Architecture of VGG-19.** Adapted from Simonyan and Zisserman (2015), Table 1. 43
2.3 **Model parameters and hyperparameters for MNIST and CIFAR-10.** β is fixed at 1 so that the KL has the same weight as the reconstruction loss; Z is fixed to have the same value as the number of classes in the datasets. 44
2.4 **Model parameters and hyperparameters for the custom dataset.** The values of the hyperparameters β and Z were chosen by performing a K-fold crossvalidation analysis, while the parameters were manually adjusted. 44
2.5 **Model parameters and hyperparameters for NSD.** The values of the hyperparameters β and Z were chosen by performing a K-fold crossvalidation analysis; α value was adapted for standard interpolation pipeline; the parameters were manually adjusted. 45
2.6 **Model parameters and hyperparameters for DTD in the VAE training experiment.** The values β was manually set to be very small to purposely underweight the variational rule; Z dimensionality was kept equal to the number of semantic classes of DTD. 45

3.1 **Convolutional Neural Network architectures grouped by family.** . 57
3.2 **Correlation between texture representation quality and Brain-Score.** Texture representation quality is expressed as MI between the clusters in the Gram representation of textures and the ground-truth labels in DTD, and the quality of the networks as models of the ventral visual stream is measured by Brain-Score. For more details on Brain-Score see Section 3.3.7. 61
3.3 **Textures generated with the Gatys algorithm** applied to a subsample of CNNs. 63
3.4 **Extracted layers from each CNN** and relative Gram matrix dimensionality. The names refer to the terminology of the models' versions in Torchvision. For models whose layers are not associated with a name, we report the absolute index. 67
3.5 **(Continued from previous page)** 68
3.6 **Summary of the values from Brain-Score** benchmarks that we selected for our study (columns), for the 12 CNNs from our pool (rows). In the second column and second row, for models and benchmarks respectively, we report the .csv coordinates of the leaderboard. Our Brain-Score values were exported from the leaderboard website on September 8th 2025. 69
3.7 **Textures generated with the Gatys algorithm** with the 12 CNNs of choice (rows). The 4 original images (top row) belong to 4 different classes in DTD (columns: *blotchy; striped; matted; scaly*). The last column represents one of the images originally used by Gatys et al., which we synthesized as a reference. 71
3.8 **(Continued from previous page)** 72

4.1 **Summary of model's details.** Consistently with previous studies, we used VGG-19_bn. Indeed, batch normalization layers allow a stable feature representations of textures (de Paolis et al., 2026; Gatys et al., 2015). We extracted every batch normalization layer after any convolutional layer, for a total of 16 layers. 103
4.2 **Summary of table humans-models correlations for Section 4.5, Figure 4.4b**. 104

CHAPTER 1

Introduction

This thesis presents a series of interconnected research studies on visual textures, an important but still understudied topic of vision neuroscience. Questions about human vision are often framed in terms of conscious recognition: how do we understand what we see? In practice, this has usually meant asking how we recognize objects, as if seeing the world was only a matter of identifying the things around us. Much less attention has been given to a complementary question: how do we see what we cannot easily name or understand? Textures belong precisely to this niche area of vision. They are everywhere in natural scenes, they strongly shape visual experience, and yet they haven't been recognized as separated phenomena within the object-centered paradigm.

The object recognition framework has been extraordinarily productive. It has guided the discovery the brain processes underlying the perception of visual stimuli, the representation of objects, and it has shaped the development of computational models of vision, influencing the release of successful modern computer vision systems. The object recognition framework has also brought together diverse research disciplines such as neuroscience, psychophysics, computer science, linguistics, and others around a common goal: understanding how visual information makes sense. However, the same paradigm also leaves open a fundamental question: if vision is not only about recognizing objects, what else does the visual system represent? Parallel to studies on object recognition, other lines of research in vision psychophysics began to focus on aspects of vision that often occur outside awareness. In this context, visual textures are especially useful: they make it possible to study perceptual phenomena and implicit image processing without relying on object identity. Synthetic textures generated from computational models of low-level image features, were used in psychophysics and physiology experiments to study which information the visual system extracts from visual scenes. Within this fertile area of research, textures gradually emerged not only as convenient stimuli, but as acknowledged objects of study themselves; in fact, they provide experimental access to dimensions of vision that remain partly hidden when scientific questions only focus on objects, categories, and recognition.

My contribution to the study of visual textures lies at the intersection among computer vision, vision neuroscience, and data science. In the early stages of my PhD, I developed a framework to study the perceptual properties of synthetic visual textures. Over time I investigated a more general research question: what kind of visual information do textures carry, and how is this information represented? Following this lead, I researched beyond the visual properties of texture images and focused on the perceptual and semantic representations formed by humans and computational models of vision. I brought together synthetic and naturalistic textures into the same experimental pipeline, considered the object recognition paradigm as a theoretical reference point, and questioned whether this paradigm is sufficient to capture the nature of textures and their representational properties. Overall, this thesis presents a computational approach to the study of vision through the lens of visual textures. By doing so, it aims at clarifying what textures reveal about internal representation formed by humans and vision models, and at outlining future directions with potential implications for the theoretical foundations of vision science.

In the following sections of the Introduction I will provide useful background notions to understand each research project. In Chapter 2 A Variational Autoencoder for texture

synthesis I will discuss the development and training of a Variational Autoenconder with a convolutional backbone to generate high-quality, naturalistic-looking artificial textures. The model synthesizes artificial textures by optimizing an initially randomized image to match the texture contained in an input naturalistic image. I will describe how this model failed to produce the expected results despite attempting several training regimes and different datasets, and despite isolating a number potential sources of error. Nevertheless, I will discuss how the negative result lead us to examine new open questions around defining efficient representations of textures. Chapter 3 Perceptual misalignment of textures representations in convolutional neural networks is dedicated to the analysis of a well known algorithm to extract texture representations from naturalistic images, invented by Gatys et al. (2015) and originally implemented with a specific Convolutional Neural Network. I implemented the algorithm with a large pool of Convolutional Neural Networks trained for object classification, and asked if the texture representations extracted with the Gatys' algorithm align with traditional measurements of human-likeness in vision models. Finally, in Chapter 4 Texture representations in deep vision models: comparing Convolutional Neural Networks, Vision Transformers, and human perception I will introduce the usage of Vision Transformers as models to extract reliable texture representations, and will compare such representations with human performance quantified with a behavioral task. I will show that Vision Transformers can produce meaningful texture representations which are also aligned to human perception. To conclude, I will summarize the main findings, and outline future directions that could potentially impact the theoretical foundations of computational vision and texture modeling, which include studying texture perception separately from object recognition and through usage of Vision Transformers rather than classic CNN models.

1.1 The ventral visual stream

Vision is a complex process that involves several stages, from the eyes to many brain areas. The eyes receive inputs from the external environment and elaborate raw visual information by means of two cells: rods and cones. Rods and cones are neurons specialized for photoreception and populate the retina. Cones are concentrated in the central part of the retina, the fovea, and respond to luminous stimuli of different wavelengths; rods are concentrated in the parafovea, the peripheral part of the retina, and respond to to motion and absence of light. As a consequence to receiving visual stimuli, rods and cones transmit electrical signals via the optic nerve, through the Lateral Geniculate Nucleus, and to the rest of the brain; here, the signal reaches to the primary visual cortex, located in the two occipital lobes. In primates, the primary visual cortex, known as striate cortex or as V1, is the input stage of the visual cortex, which is a large brain region devoted to the processing of visual stimuli. After V1, visual information crosses to two different pathways of the visual stream, called dorsal visual stream and ventral visual stream. The former elaborates visual stimuli in motion and expands towards the dorsal side of the occipital lobes; the latter processes static visual scenes and it goes deeper into the temporal lobes. I will focus on the ventral visual stream as a biological foundation basis of my research, which comprehends the following extrastriate areas: V2, V4, and inferotemporal cortex (IT).

The ventral visual stream is organized as a hierarchical process of visual stimuli in the areas mentioned above (Desimone and Schein, 1987). V1, or striate cortex, is the first area to receive signal from the retina, and it selectively responds to stimulus orientation and luminance. A series of influential studies on the mammalian striate cortex were published

by Hubel and Wiesel. They studied the receptive field and the firing patterns of V1 neurons in lightly anesthetized cats (Hubel and Wiesel, 1959); mapped the neuronal organization and retinotopy of the cats' striate cortex (Hubel and Wiesel, 1962); and eventually they studied primates' V1 (Hubel and Wiesel, 1968). Specifically, in Hubel and Wiesel (1962), Hubel and Wiesel recorded the firing activity of V1 neurons and described a small neuronal receptive field size. Using a bar of light projected onto a screen and manipulating its orientation and position they discovered two types of fundamental cells: simple cells, which selectively fire for stimulus shape, orientation, and position confined to the size of the local receptive field of the neuronal population; complex cells which receive inputs from simple cells, respond to the same stimulus features as simple cells, but they generalize the positional response to larger areas other than local neighborhoods. A few years later, Hubel and Wiesel (1968) extended their study of the striate cortex to primates with in vivo electrophysiology, thus providing a model linked and relevant to the human visual system. To this date, the seminal work by Hubel and Wiesel remains a fundamental element in the modern understanding of the mammalian visual cortex.

The output signal of V1 flows into V2. Studies on primates' V2 revealed that neurons in V2 have an approximately double sized receptive field than that of V1 neurons (Gattass et al., 1981). V2 neurons were also found to respond to regular gratings, contour shapes (Hegdé and Van Essen, 2000), angle orientation (Ito and Komatsu, 2004), combinations of different orientations, and simple textures (Anzai et al., 2007). A very small fraction of V2 neurons was also associated to the processing of stimuli with different color wavelengths. Schmid et al. (2009) propose a functional distinction of neurons populating V2: "sustained neurons", which elaborate the same stimuli as in V1 with a temporal decay; "transient neurons" which are sensitive to temporal dynamics and respond to higher level visual features such as texture, depth, simple shapes and lines. Also Willmore et al. (2010) describe a heterogeneous neuronal population in V2. They find that half of the neurons respond to similar spatial tuning to those of V1, and the the other half encodes more complex image features such as those occurring in natural visual scenes. In essence, the role of V2 is complementary and additive to that of V1: it contributes to a higher level of abstraction of image processing by integrating and differentiating spatially and temporally determined stimuli.

Following V2, V4 builds on the hierarchical visual process by receiving feedforward, direct inputs from both V1 and V2, and by sending feedback outputs (Pasupathy et al., 2020). Desimone and Schein (1987) find that most cells in V4 are highly selective for stimulus form, and Pasupathy and Connor (1999) find that V4 processes information about contour and boundary features. Similarly to V2, V4 selectively responds to orientation, spatial frequency and more strongly to color, as well as to more complex object features such as structured shapes. The authors also show that the receptive field of the neuronal population in V4 is 4 to 7 times larger than those of V1 neurons. In an interesting review, (Pasupathy et al., 2020) argue that V4 is likely to be involved in the construction of objects representations taking place in the Inferior Temporal cortex by selectively responding to contour features and by suppressing representations of textures received through direct connectivity from V1 and V2. This consideration is particularly relevant to understand where and how textures representations are formed in the ventral visual stream, and to which extent they contribute to or stray from the representations of objects. To summarize, the role of V4 can be described as the processing of visual information of texture, shape and color as a step toward complex object recognition.

The final stage of the ventral visual stream is the Inferotemporal Cortex (IT), located in the inferior portion of the temporal lobes, and putatively shares an associative net-

work with other sensorimotor cortices to contribute to the formation of semantic memory (Patterson et al., 2007). IT selectively respond to and representations of objects, body parts, and faces by receiving inputs from and sending feedback to V1, V2 and V4. Tanaka (1996) shows that IT neurons selectively respond to various moderately complex object features arranged as a continuous map in a complex feature space, and Logothetis and Sheinberg (1996) argue that in IT there is specific system for general object recognition, while there is an integration of multiple representations based on different image transformations. Later, Rust and DiCarlo (2010) find that V4 and IT share similar representational properties of objects, but the effective accomplishment of such task in IT is mainly due to the difference of the receptive field size. This result further supports the importance of the size increase of neurons' receptive field across the ventral visual stream to achieve object recognition. Finally, (DiCarlo et al., 2012) introduce the concept of "core object recognition", which is an immediate and effortless mechanisms invariant to image transformations, that preserves object identity across real-world variations.

1.2 Computational models of the ventral visual stream

In parallel to the biological studies understanding the role of the ventral visual stream, its functioning and its organization, the scientific community committed to deepen this knowledge by proposing a number of computational models. In the 1960s and 70s, several studies made an effort to mathematically characterize the functioning the cells present in V1. The summation of the receptive fields of simple cells is described as a linear process related to the contrast or luminance of the stimulus (Enroth-Cugell and Robson, 1966; Movshon et al., 1978a), while the summation of complex cells' receptive fields is mostly nonlinear: in fact, the receptive fields of complex cells seem to interact nonlinearly with each other by means of smaller subunits (Movshon et al., 1978b). In an important review, Carandini et al. (2005) report that initially single linear filters were preferred to model V1 neurons, but they were later substituted by nonlinear models and multiple nonlinear filters. The authors thus advance the hypothesis that just linear or just nonlinear models could fall short in fitting the description of early and middle vision stages. Notably, inspired by Hubel and Wiesel's work, Fukushima (1980) developed a pattern recognition model named Neocognitron. The Neocognitron is designed and tested after Hubel and Wiesel discoveries on the functioning of V1 (Hubel and Wiesel, 1962, 1968). The Neocognitron is composed by several modules equipped with with S-cells and C-cells, named respectively after simple cells and complex cells. S-cells and C-cells were designed to model their biological homologue, but their activity is repeated over several layers, making the Neocognitron a hierarchical model that mimics not just V1, but the whole ventral visual stream.

As reported in Kuypers et al. (1965), there are feedforward connections between V1 and V2. They describe two distinct types of transformations happening between the visual two areas, which are particularly relevant for texture perception. The first, called differentiation, enables V2 to extract texture boundaries; integration builds on the boundary information extracted in V1. The authors explain that the type of responses detected in V2 from continuous and discrete stimuli – such as gratings – are well represented by nonlinear dynamics, and argued that, while feedforward networks can model temporal differentiation and feedforward inputs from V1, a more complex model is required to describe non-spatial differentiation and feedback connections with V1 and possibly other areas. Freeman and Simoncelli (2011) later developed a mathematical model that describes the increase of neuronal populations receptive fields in V2, approximately double

in size than those in V1, as a function of stimulus eccentricity. They generate synthetic images that look identical to the original in the central fixation point with a parametric method (Portilla and Simoncelli, 2000a), but with distorted peripheral features. Images whose distortion was determined by a pooling size parameter that matched the receptive field size in V2, were indeed found to correlate V2 responses. Since V2 was linked to texture perception (see in the previous paragraph Ito and Komatsu (2004); Schmid et al. (2009)), Ziemba et al. (2016) later developed a framework to test directly the link between V2 and sensitivity to visual textures. Using the same method as in Freeman and Simoncelli (2011); Portilla and Simoncelli (2000a), they generate synthetic samples of naturalistic visual textures characterized by different types of patterns that look identical to the original, but their statistics of local textures ensembles are tightly controlled to match the sensitivity of V2 neurons. In fact, they find that V2 neuronal responses match such statistics, while having tolerance for other image variations. The work by Ziemba et al. (2016) established a direct link between texture perception and the functioning of V2, marking an important discovery in the area of vision neuroscience that ties computational modeling of textures and experimental physiology.

About V4, one influential computational model is the Spectral Receptive Field (SRF), designed from observations about the selectivity of V4 neurons for complex visual patterns, such as contours, gratings and edges with various orientations (David et al., 2006). SRF essentially depicts the shape tuning behavior of V4 neurons by means of a second-order nonlinear model that describes tuning in terms of stimulus orientation and spatial frequency power spectrum. The findings made with SRF are consistent with previous observations that linearity can't explain higher-order visual phenomena in the later stages of the ventral visual stream, beyond V1 (Carandini et al., 2005). Later, Pasupathy and Connor (2001) study the response of single V4 cells to position-specific curved boundaries in the framework of object recognition. They employ closed-shape visual stimuli featuring convex or concave edges and quantify the tuning of V4 cells to such stimuli by using a set of complex Gaussian functions determined on a curvature $\times$ position domain. They fit the model with in vivo recorded neural response and developed a method that is now considered an important mathematical framework that further characterizes the complex and specific functions of V4 neuronal population. Later, Cadieu et al. (2007) published an explicit computational model of shape selectivity in V4 to address how this function contributes to object recognition. In particular, they test how shape sensitivity emerges from the direct connection between V1 and V4. The model has parameters that are specific to V4 cells, as evidenced in Pasupathy and Connor (2001), the architecture is hierarchical and feedforward, and the whole structure mimics the functioning of the ventral visual stream from V1 to V4 motivated by previous experimental evidence of the functioning of each visual stream area.

The last stage of the ventral visual stream is the inferior temporal cortex (IT), where object recognition takes place. In an influential review, Riesenhuber and Poggio (2000) underline that object recognition is traditionally considered both an identification and a categorization problem. Identification stands for the accurate discrimination of objects, while categorization stands for the ability to group together sets of objects. This framework later led to the development of computer vision systems that solve either one of them. However the authors explain how, in biological vision, object recognition happens in IT after the hierarchical stages of image processing in the ventral visual stream: they highlight how IT is capable of recognizing and categorizing object easily and quickly ($\simeq 150ms$) and how it produces representations that are robust to image variations. Consequently, Riesenhuber and Poggio (2000) support the efficacy of feedforward and hierarchical mod-

els such as Neocognitron (Fukushima, 1980) and the subsequent HMAX (Riesenhuber and Poggio, 2000), that describes more reliably the feedforward connections between V1, extrastriate visual areas, and IT which contribute to the realization of invariant object recognition. Anselmi and Poggio (2014) review and discuss evidence that the main role of the feedforward path of the ventral visual stream is to encode object invariant representations of images. In particular, the authors argue that within the first $\simeq 100ms$, areas V1, V2 and V4 collectively and progressively form object representations that are robust to variations such as translation and scaling; in IT, another stage of image processing provides representations that are apt for object classification and recognition, based on the identity transformation invariance learned in the earlier stages of the ventral visual stream.

Early theoretical models of biological object recognition, in fact, either missed the hierarchical structure, or the feedforward connectivity. In the Recognition-by-Components (RBC) theory, Biederman (1987) propose that object recognition happens thanks to the spatial overlay of the so-called "geons", volumetric primitive entities linked by spatial relations. Later Mel (1997) propose SEEMORE, one of the earliest biologically inspired computational models of object recognition. SEEMORE is a hierarchical and feedforward system that integrates shape, texture and color cues through a series of nonlinear operations. More recently, Riesenhuber and Poggio (1999) published HMAX. As mentioned above, HMAX is considered a further, important step to a successful and biologically reliable model of object recognition in the visual stream. HMAX, in fact, builds its architecture on the precursor Neocognitron (Fukushima, 1980), discussed earlier. Also in the case of HMAX, the seminal work by Hubel and Wiesel (1962) constitutes the main biological foundation: HMAX has are two types of operating units: the first type largely overlaps with the functioning of the Neocognitron's S-cells, and the second type recalls C-cells. The function of the first type of cells is similar to S-cells, while the second type replace C-cells with pooling operations over the activity of the first type of cells. In HMAX, the activity of these two types of neurons is built in a number of stacked layers that constitute a feedforward hierarchy, becoming one influential type of neural model of the ventral visual stream. As a secondary consequence, HMAX was also shown to capture the response time span of deeper cortices in the ventral visual stream, enabling its biological plausibility.

DiCarlo and Cox (2007) later brought forward, in a popular review, the idea that biological object representations lie in high dimensional, tangled manifolds generated by hierarchical processing in the ventral visual stream. In this paper, the authors introduce a new crucial perspective to solve the problem of object recognition, both biologically and computationally. They propose a framework in which object recognition is fundamentally a geometric problem: the representations of objects that support both object identification and categorization are fundamentally complex, because they require many neuronal processes from the retina till IT. The authors suggest that this problem, considered from a geometrical perspective, can be represented as thousands of retinal images of objects and scenes organized as intertwined manifolds. They suggest that the solution is to identify those brain processes that disentangle object representations, such as manifold flattening. This solution, in turn, could serve as a guidance to build reliable object recognition models robust to identity-preserving transformations and fast as the brain. In a later review, DiCarlo et al. (2012) introduce the concept of "core object recognition", formalizing the ideas introduced in previous research works (DiCarlo and Cox, 2007). They explain that object recognition in primates happens in a short interval within the average $300ms$ time employed to explore a visual scene, i.e. $150ms$ (Thorpe

et al., 1996). The authors consider this time frame and the recognition that results in this short process as the core mechanism underpinning object recognition. They once again underline that object recognition is a complex process that involves complex overlapping geometrical representations in high dimensions, but they argue that reducing the problem to core object recognition simplifies greatly the search for a solution. The problem of object recognition showed interesting advances and the rise of new theoretical hypotheses, such as those in support of a functional overlap between mammalian visual cortex and artificial neural networks for vision, which led to the one of the most influential studies that mapped the representations of objects formed in IT to those formed by a large pool of computational vision models, spanning early models, such as HMAX, to the modern convolutional neural networks. Using Representational Similarity Analysis, Kriegeskorte et al. (2008) showed that supervised models achieved the closest representations of objects to those of human IT gathered with functional MRI, with the top performance reached by the convolutional neural network (Khaligh-Razavi and Kriegeskorte, 2014).

1.3 Artificial vision

In the early 2010's the computer science research community adopted the usage of computer systems inspired by the challenge of biological object recognition for improving automated object classification in the field of engineering. Among the most famous vision models there are Convolutional Neural Networks (CNN), developed in the field of computer vision to classify images. CNNs were partially inspired by precursors such as Neocognitron and HMAX (Lindsay, 2021). Computer vision systems were usually tested with challenges such as ILSVRC (ImageNet Large Scale Visual Recognition Challenge (Deng et al., 2009)), which consisted in evaluating the performance of the model with an unreleased database of images to test the ability of candidate models to perform automatic object classification. The advent of CNNs deeply impacted the field of computer vision, and later also the neuroscience of vision, from which it initially took inspiration.

CNNs became very popular with the release of AlexNet (Krizhevsky et al., 2012), which won the ImageNet competition with a significantly higher accuracy than other vision models. AlexNet is a large CNN equipped with 5 convolutional layers, some of which are followed by max-pooling layers, and it ends with a softmax classification head with 1000 units, the same number as the categories in ImageNet. AlexNet is a feedforward neural network trained with the backpropagation algorithm based on gradient descent (Rumelhart et al., 1986), and it uses dropout as a regularization method (Srivastava et al., 2014). AlexNet performed significantly better than the rest of the models in ILSVRC-2010 and entered the competition in the edition of the challenge ILSVRC-2012 (Russakovsky et al., 2015), surpassing the best second model by 11% in the top-5 test error rate, with a score of 15,3%. After AlexNet, the computer vision community released many CNNs trained on large amounts of images for object classification using backpropagation. These models vary in size and architecture, but overall the convolutions plus pooling operations stacked in different layers create image representations compiled into feature maps that become increasingly more complex (Güçlü and van Gerven, 2015). CNNs are designed and trained to gradually process local image features with a level of detail that increases as a function of depth. In the latest layers, associated with the ability to extract relevant semantic information from the input images, the level of detail peaks. This process serves to achieve object classification, the main objective of CNNs first exemplified by AlexNet. The deep hierarchical architecture was designed to provide an efficient method to infer semantic attributes from images of objects. The deep hierarchy gradually transforms the

input image to enable the extraction of semantic information defined as the class label to which the image is associated. The convolutional layers, that perform convolution operations over different stages of the image processing, feature receptive fields (RFs), the computational version of the biological receptive fields of neurons in the visual cortex. In characterizing CNNs, RFs serve as local image detectors: their size increases with depth, and by the output head the RFs act as global image features detectors. Finally, pooling operations serve as a subsample method to decrease information redundancy among neighboring neurons within a certain layer. Pooling operations compute the mean over a number of local RFs outputs of the convolutional layers, and are crucial to keep the image representations tolerant to identity-preserving object transformations.

Due to these characteristics – deep hierarchy, convolutions, pooling – were indirectly linked to the biological visual system of primates and humans and its functioning: namely, both of them enable object recognition in the very last stages thanks to progressive image processing stages, simple cells – or convolutions – complex cells, – or pooling operations –, RFs size growth and a combination of linear and nonlinear operations. However, it is important to highlight that CNNs were not designed to explicitly model the biological functioning of the ventral visual stream. The main goal of the computer vision community was to improve computer vision system to perfect the task of object classification taking inspiration from the neuroscience of vision but with no ambition to provide a mechanistic explanation for object recognition in the mammalian cortex. Therefore, at this stage the mechanistic explanatory link between large scale deep neural networks and object recognition in the brain was not yet established.

After 2012, other types of deep convolutional models for image classification were published. The main goal was to improve object categorization to achieve human-like performance, or to optimize computing resources. These novel CNNs, in fact, are larger than AlexNet, with many more parameters to adjust, however in large CNNs the depth of the model can affect gradient descent, consequently backpropagation, and finally image classification performance (He et al., 2016). ResNets, along with the classical feedforward connections, were also equipped with residual connections between the first layers and later layers, in order to keep the gradient descent stable trough all the stages of image processing and avoid gradient vanishing or explosion. During the same years, two other relevant CNN architectures were developed: VGG Simonyan and Zisserman (2015) and InceptionNet (Szegedy et al., 2015). Both models were designed to improve object classification performance, but while VGG exploits a larger number of convolutional and pooling layers, InceptionNet uses inception layers, an alternative method to improve object classification performance without consuming excessive computational resources. The two VGG versions, VGG-16 with 16 convolutional layers, and VGG-19 with 19, reached the top two places in ILSVRC14 Russakovsky et al. (2015). VGG significantly improved object classification performance by steadily increasing the amount of convolutional layers before the classification head, and by varying systematically the number of features along the depth of the architecture the size of the convolutional filters along the architecture depth. Both VGG models accepts as an input images of size $I \in \mathcal{R}^{224\times224\times3}$, and transmit it to the first block of convolutional layers whose RF size is 3×3. The architecture of both VGG-16 and VGG-19 is organized as follows: 5 blocks composed by 2 or 3 (or 4, in the case of VGG-19) convolutional layers are followed by a maxpool layer. The RFs of the convolutional layer in each block become progressively larger. After the last maxpooling operation, three fully connected layers follow, and the final head outputs a probability score for classifying the images in the correct label. A synthetic overview of the architecture can be found in Table 1.1.

VGG-16	VGG-19
Input image ($224 \times 224 \times 3$)	
conv-64 conv-64	conv-64 conv-64
maxpool	
conv-128 conv-128	conv-128 conv-128
maxpool	
conv-256 conv-256 conv-256	conv-256 conv-256 conv-256
maxpool	
conv-512 conv-512 conv-512	conv-512 conv-512 conv-512
maxpool	
conv-512 conv-512 conv-512	conv-512 conv-512 conv-512 conv-512
maxpool	
fully connected 4096	
fully connected 4096	
fully connected 1000	
softmax	

Table 1.1: **Architectures of VGG-16 and VGG-19.** Adapted from Simonyan and Zisserman (2015), Table 1.

Soon the neuroscience community realized that, although not being as biologically plausible as spiking neural networks, CNNs were still able to capture some functional properties of the ventral visual stream. Consequently, scientific interest grew around the idea that CNNs could be useful for understanding information processing in the brain, as well as learning higher levels of abstraction of visual scenes representation Kriegeskorte (2015). During the same years, Khaligh-Razavi and Kriegeskorte (2014) discovered that unsupervised CNNs produce similar representations to IT. Furthermore, Yamins et al. (2014) found that the performance in object categorization tasks of a CNN correlates with its capability to predict the neural responses of IT, that the latest layers before the categorization head strongly predicts IT activity, while intermediate layers predict V4 activity. Subsequently, Cadieu et al. (2014) shows that deep CNNs rival the performance of humans in an object categorization task and that humans and deep CNNs produce similar representations of such images, while Güçlü and van Gerven (2015) discovered that the early layers of a deep CNN trained on object recognition develop image representations compatible to the image features produced in V1, commonly referred to as "Gabor filters". These findings opened to the explorations of image representations formed in CNNs, rather that asserting a direct biological comparison with the functioning of the brain. CNNs were deemed capable of extracting lower-level visual features similar to those found in the early visual cortex, and of approaching the development of complex semantic features similar to those associated to the representations of objects and scenes in IT Kriegeskorte (2015).

CNNs were therefore considered by the neuroscience community as established functional models of the ventral visual stream (Lindsay, 2021). More recently, a famous benchmark that quantifies the goodness of a large pool of vision models in modeling biological object recognition (Schrimpf et al., 2018) established that the VGGs are one of the best models available, in line with previous findings (Güçlü and van Gerven, 2015; Yamins et al., 2014).

More recently, a new kind of architecture for computer vision was developed. These models are commonly called Vision Transformers (ViT) (Dosovitskiy et al., 2021), as they are based on the attention mechanism as the first Transformer type of neural network, which was invented as an alternative to deep recurrent neural networks and deep CNNs for sequence modeling Vaswani et al. (2017). The core component of the Transformer is the self-attention mechanism, which allows each input sentence token to dynamically integrate information from all other tokens in the sequence. Each token is projected into query, key, and value representations respectively. Attention weights are computed from the similarity between queries and keys, and then are used to produce a weighted combination of the value vectors. As a result, each token acquires a context-dependent representation that captures both local and long-range relationships within the input. Vision Transformers extend the Transformer architecture to image processing by treating an image as a sequence of patches, instead of tokens. Through self-attention, each patch dynamically integrates information from all other patches in the image, allowing the model to capture both local and global relationships by attending to all possible patches combinations.

One of the earliest ViTs was developed by Dosovitskiy et al. (2021). This model is a fully attention-based model, trained and tested with ILSVRC-2012 ImageNet (Deng et al., 2009; Russakovsky et al., 2015) to perform several benchmark tasks, such as object classification. In order to attend to all the positional sequences in the images, this model was designed to take as an input an image of dimensions $I \in \mathcal{R}^{H \times W \times C}$ and to convert it to a sequence of flattened 2D image patches of fixed size $x \in \mathcal{R}^{N \times (P^2 \cdot C)}$, where H, W are the spatial dimensions of the image, C is the number of channels, P, P is the size of each image patch, and $N = HW/P^2$ is the resulting number of patches which serve as an input sequence to the ViT. A visual overview of the architecture can be found Figure 1.1.

In the same period, (Chen et al., 2020) developed a novel Transformer architecture model for image processing, commonly known as image-GPT (iGPT). iGPT is an decoder-only generative model pretrained to predict pixels instead of sentence tokens as in the original use by Vaswani et al. (2017). Interestingly, the authors also observed that iGPT forms high-quality unsupervised image representations with no prior semantic information. Furthermore, iGPT was shown to achieve competitive scores to those of CNNs in image classification tasks with several benchmark datasets, including images contained in ImageNet. Since iGPT is a generative model, the object classification task was implemented by attaching a linear classifier head to the layer that produces the best image representations.

Radford et al. (2021) developed CLIP (Contrastive Language-Image Pre-Training), a ViT that produces multimodal embeddings from both language and images inputs. CLIP has two encoders, one accepting images and the other accepting captions. While the language encoder is a transformer, the image encoder can be either a CNN or a ViT, but in this thesis I will focus on the ViT based model. By means of a contrastive loss, CLIP learns embeddings of both images and text features in a self-supervised regime guided by the language conveyed in the captions. Despite not being trained to explicitly perform object classification, CLIP can perform zero-shot transfer learning for object

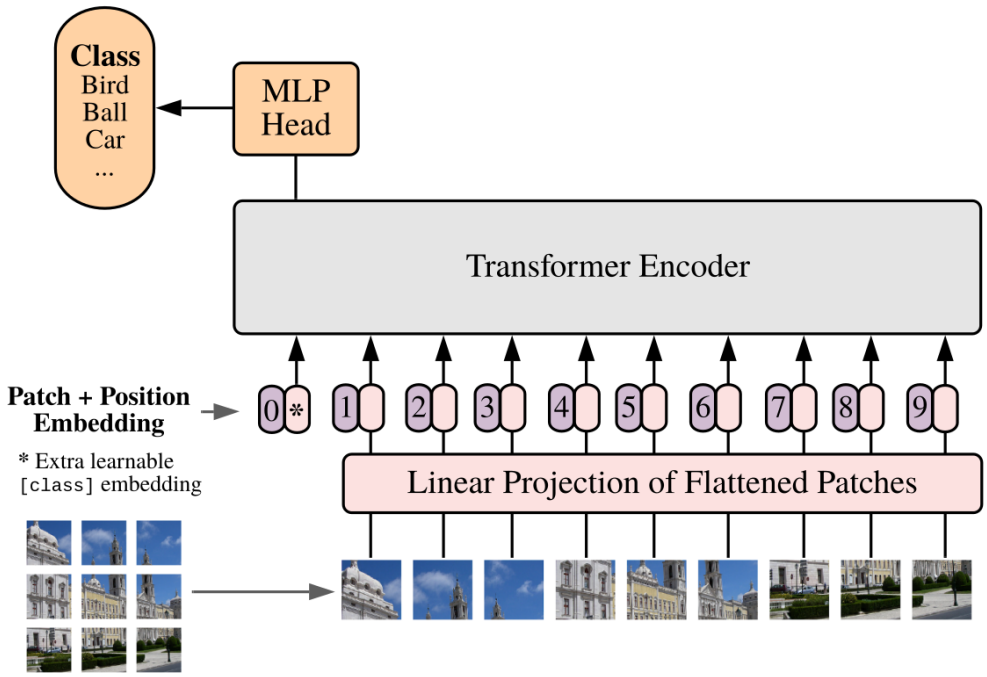


Figure 1.1: **Architecture of the Vision Transformer**, adapted from Dosovitskiy et al. (2021), Figure 1.

classification with ImageNet stimuli and labels, with accuracy scores comparable to those of CNNs. Since CLIP outputs image-text embeddings, this task was tested with a linear classifier trained and tested to associate image embeddings to caption embeddings.

Another interesting ViT model is DINO (self **DI**stillation, **NO** labels), developed by Caron et al. (2021). A second version, DINO-v2, was later developed by Oquab et al. (2024). DINO is a type of deep neural network architecture designed to study the impact of self-supervision in ViTs without labels. The authors realized that the impressive performance achievements reached by previous attention-based models could potentially be linked to the absence of label supervision in the training pipeline. They therefore explored the usage teacher-student framework with no label-guided training. They explicitly designed DINO to learn image features by distillation, i.e. via an exchange of image features from a teacher network to a student network. Most importantly, the student network input image patches are parametrized as 96×96 to catch local features, while those of the teacher network are of dimensionality 224×224, to capture global image features. The loss is a cross-entropy between the features of the student and the features of the teacher. DINO is therefore trained in a self-supervised regime, and it achieves impressive results in forming internal representations that contain scene layouts and object boundaries, as well as object classification with ImageNet and other benchmark datasets. In this case object classification was achieved by applying a k-NN classifier on raw images.

In DINO-v2, Oquab et al. (2024) extend the DINO framework to explore its ability to learn general-purpose image features, both at the image level and at the pixel level, with a training dataset based on a large quantity of curated data with no text supervision. DINO-v2 was explicitly trained on a very large amount of curated images of objects, since it scales very well with the number of data. Its performance in several tasks such as object recognition on ImageNet with the k-NN algorithm, and object segmentation, surpass the performance of DINO and extends to video segmentation, image retrieval, and semantic

segmentation in the image features.

These ViT models showed comparable if not superior performance to the state-of-the-art CNNs in several benchmark tasks, including object recognition on ImageNet data. The main difference with CNNs is that none of the ViTs described above was trained for the specific goal of object classification, and some of them were also completely naive with respect to the semantic information in the labels because of an unsupervised training regime (Chen et al., 2020; Dosovitskiy et al., 2021). They also need more data than those contained in ImageNet, because of the presence of some biases in case of insufficient data samples Dosovitskiy et al. (2021). Consequently, their training datasets are often augmented with images collected from the internet and filtered to match those of ImageNet. Thanks to these interesting characteristics and to the computational scalability, ViTs rapidly became popular for solving different vision tasks in the computer vision community (for historical reviews, see Han et al. (2023); Khan et al. (2022); for an up-to-date survey, see Waseem et al. (2026)).

1.4 Visual textures

One of the founders of the research in visual textures was Bela Julesz, who published several foundational papers about pattern discrimination, in which he combined mathematical modeling and psychophysics experiments. Julesz intended to study the human perception processes that do not involve "conscious complex recognition and comparison process" (Julesz, 1962). To achieve this goal, he designed experiments where he exposed human participants to two visual fields displaying different visual statistics and looked for a "more primitive and spontaneous process which organizes the two fields in two different entities" (Julesz, 1962). For each experiment, Julesz generated the visual stimuli artificially by controlling the statistics displayed by the two patterns present in the same image with stochastic processes and local transformations. In particular, he assessed the ability to discriminate patterns based on the difference in: first-order statistics, second-order statistics, third-order statistics, and higher-order statistics. He found that participants could discriminate easily patterns that varied in first and second-order statistics, such as black and white dots and local striped structures. He also observed that patterns differing in third or higher-order statistics remain undetected by the subjects. He concluded that low order statistics produce local clusters of similar elements that are easily detectable by the human visual system, leading to the famous "Julesz's conjecture": visual textures with the same first and second-order statistics might be indistinguishable.

In a follow-up paper, Julesz et al. (1973) explicitly investigated the discriminability among textures with different third-order statistics or different higher-order statistics. For this goal the author perfected the mathematical method adopted earlier to generate texture pairs with different higher-order statistics but identical first and second-order statistics. The authors repeated the same series of experiments as in Julesz (1962), and observed that by fixing identical first and second-order statistics but varying higher-order statistics results in the human inability to discriminate textures. This finding supports and corroborates the Julesz's conjecture, and leads to the proposal of a model that pairs the evidence gathered with new psychophysics experiments combined with the neurophysiology findings by Hubel and Wiesel (Hubel and Wiesel, 1962). Julesz et al. (1973) propose a model of the visual cortex model equipped with local features extractor and a global detector: local features extractors detect simple image features commonly found in visual textures such as dots and lines, given size and orientation parameters; then, the output of such local extractors are passed to and evaluated by global detectors, which

compute first and second-order statistics.

Later, Julesz published another influential paper in the theory of vision science (Julesz, 1981). In this article Julesz gathers experimental evidence that introduces to the notion of "textons", the low-order, minimal local features of images that allow texture discrimination during visual processing. The ontology of textons, intended as psychophysical entities constituting the basic features of textures, supports pre-attentive perception of visual textures, which is referred to as the effortless and instantaneous detection of differences in textures whose textons are different. Julesz also proposes a novel version of his original conjecture, because he disproved the first original formulation. The "modified Julesz's conjecture" states that "the pre-attentive textural system cannot globally compute third or higher-order statistics" (Julesz, 1981). In fact, Caelli and Julesz (1978) discovered the existence of novel second-order image features whose differences in textures yielded to strong discriminability in human participants, followed by other several new second-order texture classes which lead to pre-attentive discrimination. In conclusion, in the Julesz (1981) paper, Julesz refutes the first version of his conjecture, and states that pre-attentive texture discrimination is either based on the difference among basic texture elements, the textons, or it is based on the first-order statistics of the patterns. According to him, he second version holds because in this study he collected evidence that the pre-attentive textural system cannot compute statistical parameters higher than second order.

Later, Malik and Perona (1990) propose a biologically motivated mathematical model of pre-attentive texture discrimination based on early visual processing. The authors explore the employment of linear filters to model texture discrimination, and ask wether linearity alone can account for human texture perception. The model reproduces the rank ordering of human discriminability measurements across several texture pairs, but they conclude that texture separation cannot be explained by purely linear filtering alone: rectification, local competition between channels, and spatial pooling are essential.

About 10 years layer, Portilla and Simoncelli (2000a) propose another influential parametric model of visual textures based on joint statistics measured in a complex steerable pyramid representation. They generate textures by decomposing an input image into oriented bands at multiple spatial scales. The model includes pixel-level marginal statistics such as mean, variance, skewness, kurtosis, and intensity range, and it includes autocorrelations of low-pass images to capture periodicity, dominant spatial frequencies, and long-range structure. Correlations between wavelet-coefficient magnitudes across nearby positions, orientations, and scales are used to represent edges, contours, and other organized features. Texture synthesis starts from Gaussian white noise and iteratively adjusts the input image so that each group of measured statistics matches those of an exemplar. The model reproduces a wide range of artificial, periodic, and natural textures, including classic counterexamples to simple low-order versions of the Julesz conjecture.

A few years later, Freeman et al. (2013) propose a model of mid-level ventral vision based on the findings observed in the texture model by Portilla and Simoncelli (2000a). In this model filter responses similar to those of V1 are combined through local correlations and pooled over regions whose size grows linearly with eccentricity. The authors synthesize visual metamers, physically different images that produce the same model responses of the original images, and test whether human observers can distinguish them. They find that for human participant, metamers and images become indistinguishable when the model value matches receptive field size found in area V2.

Victor and Conte (2012) later developed a new controlled framework for investigating how human vision responds to local image statistics that are normally found in natural

scenes. They generate binary textures defined by black and white configurations within 2×2 pixel gliders. This method leads to a set of 16 possible glider configurations, which can be reparameterized in terms of 10 independent coordinates. This 10-dimensional space is defined by one luminance-bias coordinate, four pairwise-correlation coordinates, four three-point-correlation coordinates, and one four-point-correlation coordinate. This method constructs texture ensembles in which one or two coordinates are fixed and the remaining statistics are determined with maximum entropy. Victor and Conte tested human participants in a psychophysics experiment where the texture ensembles are used as visual stimuli, and the results reveal large differences in sensitivity across statistics and show that combinations of statistics can interact. The contributions are therefore both methodological and perceptual: this study provides an effective model to generate textures with different statistics in a controlled manner and it provides evidence that visual sensitivity is organized around specific combinations of local statistics rather than around the formal coordinates considered independently. A few years later Victor et al. (2017) published a review arguing that visual textures provide an effective tool for studying vision and link evidence from psychophysics to population responses of the early visual cortex. The authors formally trace this approach back to Julesz's work (Julesz, 1962, 1981; Julesz et al., 1973), thus establishing a connection with the early discoveries and theoretical work by Julesz to more recent evidence offered the neuroscience of vision.

Just a few years later, inspired by the large usage of deep neural models in computer science, Gatys et al. (2015) introduce a parametric model of natural textures based on the feature representations learned by a CNN trained for object recognition. Instead of relying on pre-determined texture statistics, the model uses the hierarchical feature space of VGG-19 and represents textures as correlations between feature maps extracted from layers at different depths. These correlations are coded in Gram matrices, which discard the exact spatial information while preserving local co-occurrences. To synthesize a texture, the target image is passed through the network to compute its Gram matrices, and a fully random image is iteratively optimized so that its own Gram matrices match the target ones. The optimization minimizes a weighted sum of squared differences between Gram matrices and updates the image pixels through backpropagation and L-BFGS. The authors observe that when layers up to pooling layer n. 4 are included, the synthesized texture resembles the original one, improving the perceptual quality of the generated images compared to Portilla and Simoncelli (2000a). Despite not possessing any specific biological relevance, the same method was successfully applied to other applied computer vision tasks such as style transfer (Gatys et al., 2016).

A series of experiments that use the texture synthesis methods described above were published in the following years. Wallis et al. (2017) evaluate how well parametric texture models reproduce the appearance of natural textures for human observers. They compare the deep CNN model introduced by Gatys et al. (2015), the Portilla and Simoncelli (2000a) model, and a variant of the Gatys et al. (2015) model that additionally matches the image power spectrum. In an odd-one-out task, human observers choose which of three physically different image patches had been generated by a different process, thereby testing whether synthetic and original images belonged to the same perceptual texture class. Under parafoveal condition, both the Gatys and Portilla & Simoncelli models performed very well, as the observers were unable to discriminate most synthetic textures from the corresponding originals. In the foveal condition, however, the deeper Gatys model provided substantially better perceptual matches than the Portilla & Simoncelli model. The basic Gatys model nevertheless struggled with textures containing strong long-range or periodic regularities and adding a Fourier power-spectrum constraint greatly improved

the appearance of these textures.

Recently, Hermundstad et al. (2014) introduced the usage of visual textures to investigate the plausibility of the efficient coding principle, a theoretical framework in neuroscience according to which a sensory system is tuned to the statistics of its natural visual input (Tesileanu et al., 2022). In particular, the author investigated whether efficient coding principles extends to cortical representations of complex visual features. They quantify the variability of local multipoint correlations with the Victor and Conte (2012) model and use the generated textures to test human sensitivity in a rapid texture discrimination task. They find that luminance features that are more informative about visual scenes, by virtue of being more variable, are also more visually salient.

Caramellino et al. (2021) investigated whether the efficient-coding relationship previously observed between natural-image statistics and human texture perception (Hermundstad et al., 2014) also applies to rodents. Four groups of rats were trained to discriminate white noise from binary textures defined by one, two, three, or four-point correlations generated once again by the Victor and Conte (2012) algorithm. Rat sensitivity showed a clear ordering, with one and two point statistics being the most easily detected, followed by four point correlations, while three point correlations were the least discriminable. The feature sensitivity of the rats matches both human perceptual sensitivity and the variability of the same correlations across natural images. The results suggest that efficient coding of multipoint visual statistics is conserved across species and establish rats as a suitable model for studying its neural and developmental mechanisms experimentally.

Tesileanu et al. (2020) further extend the efficient coding account of texture perception from binary images to a much richer space of grayscale statistics. They define a 66-dimensional space of local luminance correlations and quantify how strongly each statistical direction varies across natural scenes. The resulting efficient coding hypothesis predicts that two-point correlations should be the most perceptually relevant, whereas correlations of higher order should contribute little to texture segregation. Human psychophysical measurements strongly support these predictions.

In a later review paper, **?** show evidence that efficient coding applies not only in the retina and V1, but also in deeper stages of the visual system. The authors discuss how perception of visual textures in humans and rodents is adapted to the higher-order spatial statistics of natural scenes, suggesting that cortical circuits preferentially represent informative multipoint correlations. These findings extend efficient coding from relatively simple peripheral features to complex texture and object-related representations in V2 and higher areas.

Recently, Zanon et al. (2025) examined whether preferences for natural image statistics arise through visual experience or are already present at birth. Visually naive chicks were raised to spontaneously approach textures defined by the Victor and Conte (2012) algorithm, and showed preferences (or lack thereof) between texture types that are compatible with the ordering of discriminability previously found in adult humans and rats (Caramellino et al., 2021; Hermundstad et al., 2014). This suggests that adaptation to the statistical structure of natural scenes is partially supported by innate biological predispositions.

1.5 Models for visual textures synthesis

In this section I will provide an overview of the three texture synthesis algorithms that I used during my PhD projects. Their historical relevance was already discussed in the previous sections. It is worth stressing that the following texture synthesis algorithms are

inspired by the early formulation of the Julesz's conjecture (Julesz, 1962). These models in fact, synthesize textures by computing low-order correlations among image features defined with different mathematical methods, as I will explain.

Hereby I will focus on their functioning as described in the original papers, while in Chapter 2 A Variational Autoencoder for texture synthesis, Chapter 3 Perceptual misalignment of textures representations in convolutional neural networks, and Chapter 4 Texture representations in deep vision models: comparing Convolutional Neural Networks, Vision Transformers, and human perception I will explain how I adapted or replicated the original algorithm for my studies.

1.5.1 Victor and Conte

This texture synthesis framework was proposed by Victor and Conte in 2012. It effectively synthesizes binary textures by mathematically defining a set of coordinates in terms of local statistics measured over all 2×2 neighborhoods of an image (Victor and Conte, 2012). This model was developed to obtain fully synthetic textures by manipulating the texture parameters found by the authors, in order to generate textures that are easily controllable in experimental settings. Other applications include the inference of the texture parameters from a given input image, without controlling for the parameters.

The algorithm is based on the idea of parameterizing luminance patterns in image. Textures are therefore synthesized in terms of pixel correlations in local neighborhoods. A neighborhood is defined as a 2×2 pixels squared glider, and because each pixel can take one of two values, a 2×2 neighborhood admits $2^4 = 16$ possible configurations.

A direct description in terms of the probabilities of these configurations is redundant. Under the assumption of translation invariance, the local distribution can instead be parameterized by ten coordinates associated with products of one, two, three, or four pixel values, which are defined as follows, and visualized in Figure 1.2:

- luminance bias: $\{\gamma\}$;
- pairwise correlations: $\{\beta_{_}, \beta_{|}, \beta_{\backslash}, \beta_{/}\}$;
- three-point correlations: $\{\theta_{\llcorner}, \theta_{\lrcorner}, \theta_{\ulcorner}, \theta_{\urcorner}\}$;
- four-point correlation: $\{\alpha\}$.

The coordinate γ determines the relative prevalence of either black or white. The β coordinates measure pairwise pixel correlations along the horizontal, vertical, and two diagonal directions. The θ coordinates measure three-point correlations that appear as a corner of a 2×2 block. Lastly, α measures the dependency among all four pixels.

Victor and Conte construct maximum-entropy image ensembles associated with specified values of these coordinates: if one coordinate is specified, they call the generated image "texture"; if two parameters are specified, the image is called "texture mixture". In both cases, after one or two parameters are fixed, the rest of the image points are determined with maximum entropy. The maximum entropy principle is important because fixing a statistic does not uniquely define an image distribution, and, among all distributions satisfying the selected constraints to generate a specific texture, the maximum entropy distribution introduces the least additional structure, which is what they want to eliminate to overcome issues encountered in previous computational models Julesz

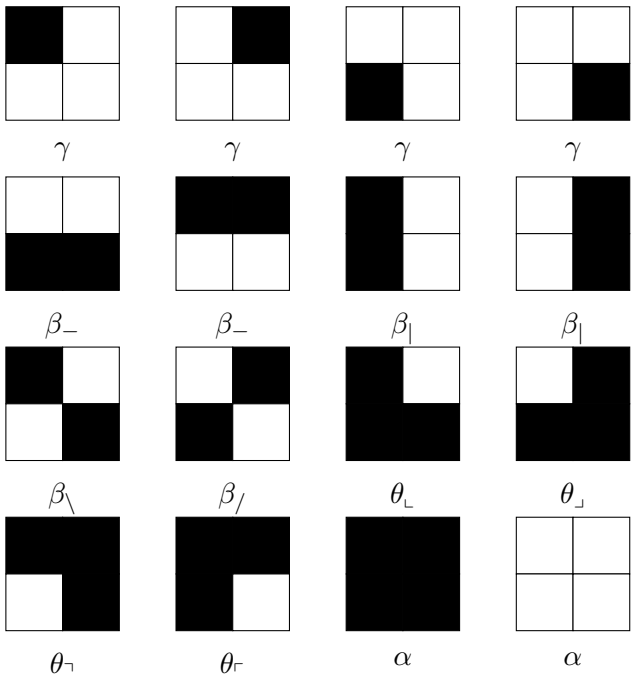


Figure 1.2: **The 16 possible binary 2×2 configurations.**

(1962). This algorithm therefore allows the perceptual contribution of individual statistics, or of controlled combinations of statistics, to be studied independently from other unconstrained regularities.

The set of parameters described above effectively defines a 10-dimensional space, where each coordinate refers to to variation along one perceptual dimension that corresponds to human sensitivity degrees (see Figure 1.3).

Textures generated with this model are binary and lack the rich structure of natural images. However, their statistical content can be controlled precisely, which makes them useful in psychophysics studies, as described in the previous section (Caramellino et al., 2021; Hermundstad et al., 2014; Zanon et al., 2025). The algorithm by Victor and Conte provides a direct way to test which local image statistics are perceptually salient in pre-attentive judgment, despite being perceptually distant from naturalistic-looking textures.

In my thesis I have used the Metex implementation of the Victor and Conte algorithm for texture synthesis by Piasini (2021). Metex provides an efficient implementation of the Victor and Conte algorithm, which, given a source image, makes it easy to infer the one or the two most dominant parameters defined in the 16 2×2 pixel neighborhoods possibilities discussed previously and to generate a synthetic texture based on them. For illustrative purposed, in Figure 1.4 show one exemplar texture defined by the most occurring coordinate that I generated from a given image, versus one texture generated by Victor and Conte (2012) from another input image.

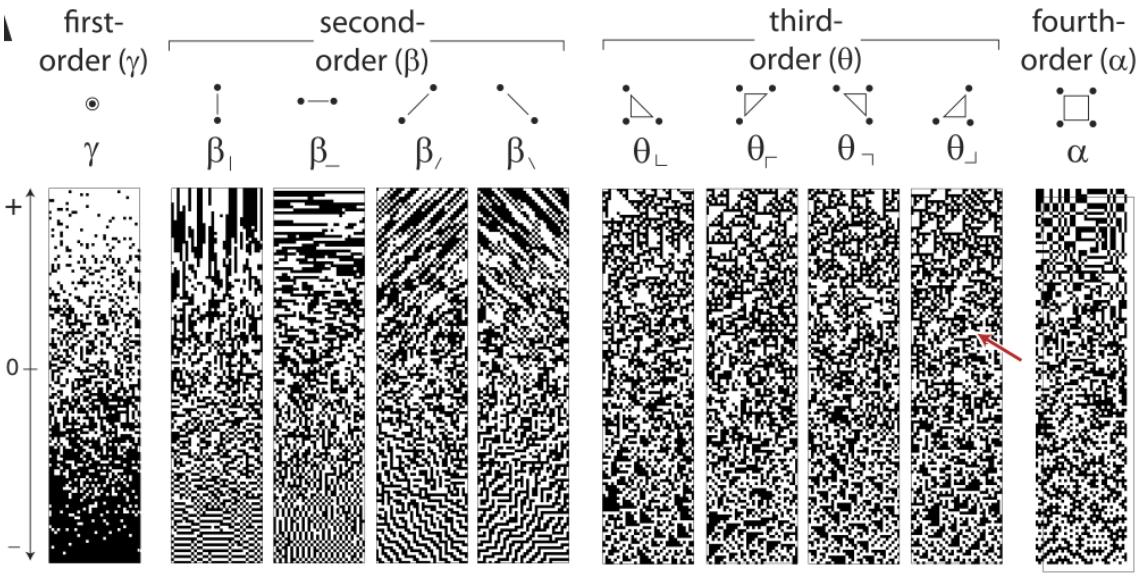
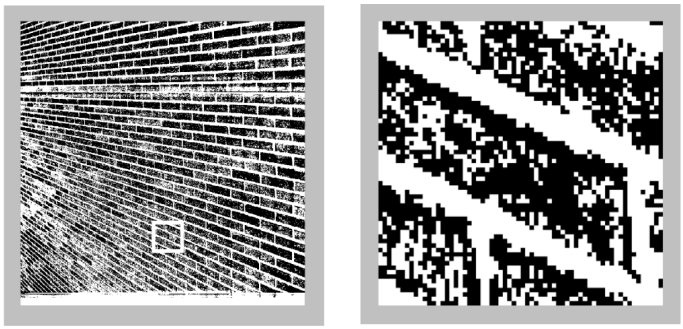
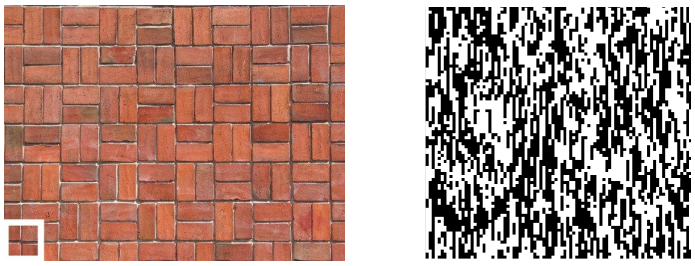


Figure 1.3: **10-dimensional parametrized space of textures** defined by the algorithm by Victor and Conte (2012). The horizontal axis ranks statistics by increasing order, starting from first-order. Adapted from Hermundstad et al. (2014), Figure 2A.



(a) Texture generated with the donut algorithm by Victor and Conte. Adapted from Victor and Conte (2012), Figure 6(a), 6(b).



(b) Texture generated with Metex from image *grid_0105.png* of the Describable Textures Dataset (Cimpoi et al., 2014).

Figure 1.4: **Comparison between similar textures** synthesized in the original paper by Victor and Conte (2012) (a), and with Metex (b).

1.5.2 Portilla and Simoncelli

The model of Portilla and Simoncelli (2000a) describes a texture through a fixed collection of summary statistics computed in both at the pixel level and a multiscale wavelet representation. Unlike the Victor and Conte model, which operates on binary 2×2 neighborhoods, the Portilla and Simoncelli model is designed to capture the spatial structure of naturalistic textures over multiple spatial scales, and to produce synthetic texture exemplars from a given naturalistic image rather than manually generate fully artificial patterns.

Given an exemplar image x, the algorithm first constructs an overcomplete steerable-pyramid decomposition. The image is separated into subbands corresponding to different spatial frequency scales and orientations. The resulting coefficients are complex valued: their magnitudes encode the local strength of oriented structures, while their phases carry information about their spatial arrangement. The statistical representation includes several complementary groups of constraints:

- marginal pixel statistics, including: mean, variance, skewness, kurtosis, and range of pixel intensities;
- the spatial autocorrelation of the image and of the low-pass residuals at different scales;
- autocorrelations of wavelet-coefficient magnitudes within each orientation and scale;
- correlations between coefficient magnitudes at different orientations within the same scale;
- correlations between coefficient magnitudes at adjacent scales;
- phase-sensitive correlations between coefficients at adjacent scales.

These statistics were chosen to capture regularities frequently present in natural textures. Marginal statistics reproduce the global intensity distribution, spatial autocorrelations constrain periodicity and characteristic length scales, and correlations among wavelet magnitudes capture the co-occurrence of edges and structures across positions, orientations, and scales. Cross-scale phase statistics additionally help preserve aligned structures such as contours.

Texture synthesis starts from an image of random noise,

$$I \sim \mathcal{N}(0, 1),$$

and proceeds through an iterative sequence of projections. At each stage, the current synthetic image is modified so that one group of its statistics moves towards the corresponding statistics of the exemplar. Since satisfying one constraint may perturb constraints imposed previously, the different projections are repeated cyclically. The procedure terminates after a fixed number of iterations or when the statistical mismatch becomes sufficiently small. The method therefore samples from an implicitly defined texture model and it does not optimize an explicit normalized probability density. An illustration of the model's algorithm is available in Figure 1.5.

The model is particularly effective for stationary and periodic textures whose appearance can be characterized by repeated or statistically homogeneous structures. It can reproduce many forms of granularity, orientation, periodicity, and multiscale organization. The synthesized texture is grayscale, somewhat naturalistic-looking, with good

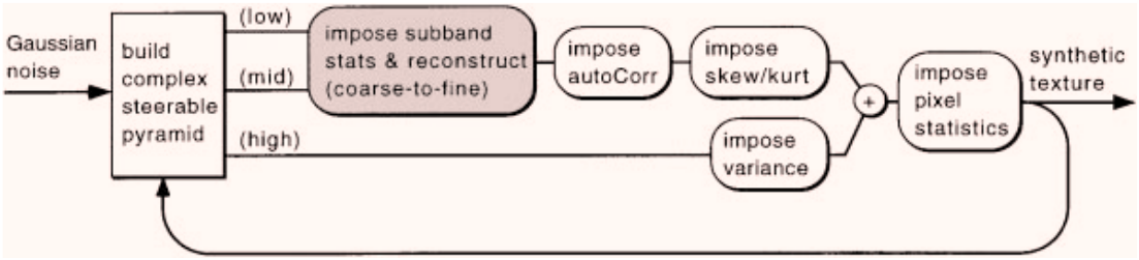
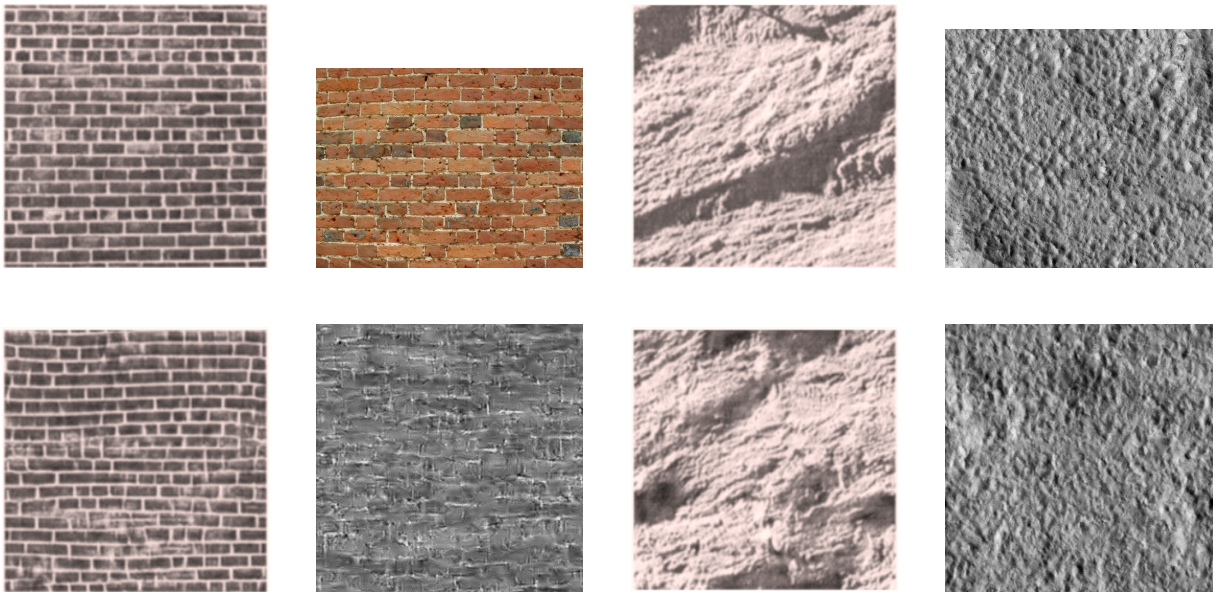


Figure 1.5: **Diagram of the recursive texture synthesis algorithm.** Adapted from Portilla and Simoncelli (2000a).

global definition which preserves the overall structure of the texture present in the input image. However, because the model only stores spatially averaged summary statistics, it generally cannot preserve the exact global arrangement of strongly nonstationary textures, neither the shape of isolated objects carrying semantic content.

In order to generate textures with the Portilla and Simoncelli model, I have used the implementation by Portilla and Simoncelli (2000b), available in the public repository textureSynth. The details of the parameterization that I used for my studies will be specifically discussed in Chapter 4 Texture Representations in Deep Vision Models: Comparing CNNs, Vision Transformers, and Human Perception. Below, in Figure 1.6 , I report a comparison between textures synthesized originally by Portilla and Simoncelli vs. similar textures that I have generated for my research:



(a) Periodic texture Portilla and Simoncelli original

(b) Periodic texture DTD

(c) Aperiodic texture Portilla and Simoncelli original

(d) Aperiodic texture DTD

Figure 1.6: **Comparison between textures synthesized** by Portilla (Figures 1.6a, 1.6c), and with the algorithm implementation by Portilla and Simoncelli (2000b) (Figures 1.6b, 1.6d), showed in the bottom row. The top row shows two input images used in the original work by Portilla and Simoncelli, and two input images from the Describable Texture Dataset (DTD) *pitted_0008.jpg*, *pitted_0025.jpg* (Cimpoi et al., 2014). Adapted from Portilla and Simoncelli (2000a) Figures 14, 15.

1.5.3 Gatys

Gatys et al. proposed that a texture can be represented by spatially averaged correlations between the feature maps of a CNN pretrained for object recognition. In contrast to the Portilla and Simoncelli model, whose summary statistics are selected manually in a wavelet representation, and in opposition to the Victor and Conte model, where the texture statistics are pre-determined as correlations among neighboring pixels, the Gatys method computes second-order statistics in the hierarchical feature space learned by a VGG-19 (Gatys et al., 2015; Simonyan and Zisserman, 2015). A visual representation of the model can be found in Figure 1.7.

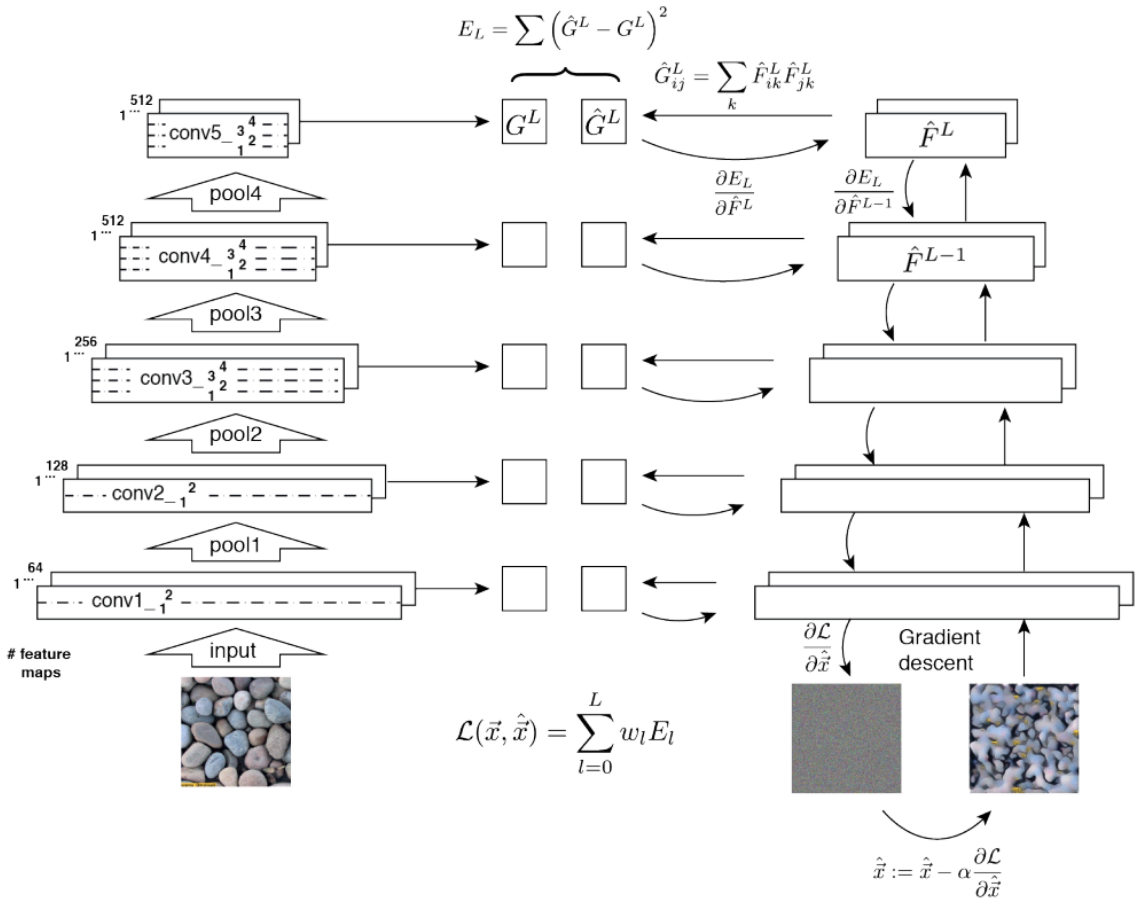


Figure 1.7: **The Gatys model.** Adapted from Figure 1 in Gatys et al. (2015).

In this framework, the texture representations are based on the Gram matrices, a 2-dimensional representation of correlations between feature maps of 5 different layers in VGG-19 that effectively discard spatial information. Let the activation of feature map i at spatial position m in layer l be denoted by F^l_{im}. If layer l contains N_l feature maps and M_l spatial positions, its Gram matrix is

$$G^l_{ij} = \sum_{m=1}^{M_l} F^l_{im} F^l_{jm}. \tag{1.1}$$

The entry G^l_{ij} measures the spatially averaged co-activation of feature maps i and j. Because the sum removes the explicit spatial index, the Gram matrix does not retain the precise location of individual features. It instead records which features tend to occur together and with what strength. This loss of spatial information makes the representation

appropriate for textures, whose appearance is often defined by stationary feature statistics than by a unique spatial configuration. An illustrative visual representation of the Gram matrices can be found in Figure 1.8.

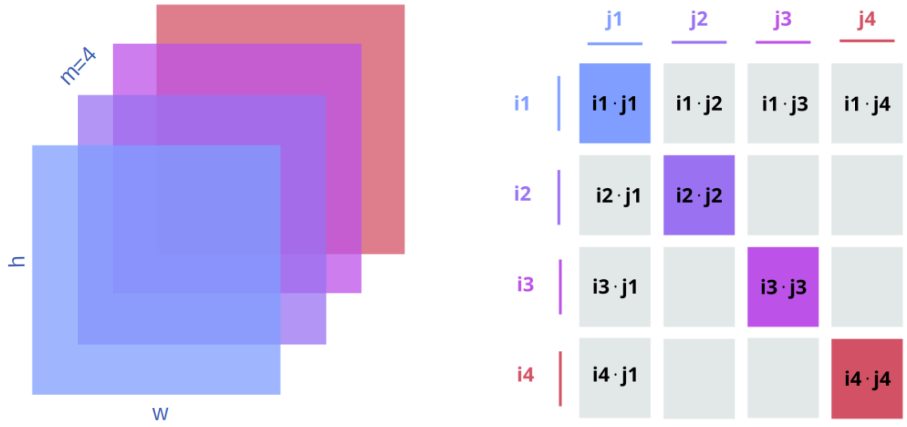


Figure 1.8: **Representation of a feature map** of dimensionality $F^{H\times W\times M}$, where $M = 4$, and its Gram matrix of dimensionality $G^{M\times M}$. The entries i, j of the Gram matrix are the dot products between all the possible combinations of the vectorized channels of the feature map.

Different CNN layers describe structure at different effective spatial scales. Early layers contain feature maps sensitive to local edges, colors, and simple orientations. Deeper layers have larger receptive fields and respond to more complex combinations of features. Matching Gram matrices at several layers therefore constrains the synthetic image simultaneously at multiple levels of scale and complexity.

Given an exemplar image x, the target Gram matrices $G^l(x)$ are first computed at the selected layers. Synthesis starts from a random image $\hat{x}_0$, whose pixels are iteratively updated to minimize the mismatch between its Gram matrices and those of the exemplar. For one layer, the loss used in the original formulation is

$$E_l(\hat{x}, x) = \frac{1}{4N_l^2 M_l^2} \left\| G^l(\hat{x}) - G^l(x) \right\|_F^2, \tag{1.2}$$

and the total texture loss is

$$\mathcal{L}_{\text{tex}}(\hat{x}, x) = \sum_{l\in\mathcal{S}} w_l E_l(\hat{x}, x), \tag{1.3}$$

where $\mathcal{S}$ is the set of selected layers and w_l controls the relative contribution of each layer. Gradients of this loss are propagated through the fixed network to the pixels of $\hat{x}$; the network weights are not updated.

Gatys et al. (2015) originally used a pretrained VGG-19 network and, for their core results, constructed the texture representation from the layers conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1 (see Figure 1.7). The original architecture did not contain batch normalization, but Gatys and colleague manually normalized the outputs of the convolutional layers (Ioffe and Szegedy, 2015) to be centered around ~ 1. In our implementation, published in de Paolis et al. (2026) and available in the public repository perceptual_misalignment, we instead used VGG-19 with batch normalization and extracted the first batch-normalization activation following the first convolution in each of its five blocks. This is therefore a modification of the original Gatys model rather than an exact

replication, as the original weights of VGG-19 pretrained on ImageNet (Deng et al., 2009; Russakovsky et al., 2015) were available in the Caffe framework for deep learning. We instead relied on the more recent PyTorch (Ansel et al., 2024; TorchVision maintainers and contributors, 2016), where a version of VGG-19 with batch normalization layers pretrained with ImageNet is available. Below, in Figure 1.9, I show a comparison of textures generated starting from the same image in the original paper by Gatys, and with my implementation. It is important to note that a direct replica of the algorithm and of the original results is not possible, as the full code and parameters for optimization are not publicly available. An in-depth discussion of the topic is available in Chapter 3 Perceptual misalignment of texture representations in convolutional neural networks.



(a) Original image (b) Gatys – LFBGS (c) de Paolis – Adam 130k (d) de Paolis – LFBGS 6k

Figure 1.9: **Textures generated artificially with the Gatys model.** Original implementation vs. my implementation. Figure 1.9a shows the original input image, a picture of pebbles; figure 1.9b shows the synthesized texture by Gatys et al. with the original implementation and the optimization algorithm LFBGS; figure 1.9c shows the synthesized texture with my implementation with Adam optimizer at 130000 steps; figure 1.9d was synthesized with my implementation and LFBGS at 6000 steps. Adapted from Gatys et al. (2015) and de Paolis et al. (2026).

Bibliography

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 929–947. ACM, 2024. doi: 10.1145/3620665.3640366.

Fabio Anselmi and Tommaso Poggio. Mit dspace publication record. https://dspace.mit.edu/entities/publication/8e222c2d-4a02-45e6-8036-2da8285bd983, 2014.

Akiyuki Anzai, Xinmiao Peng, and David C. Van Essen. Neurons in monkey visual area V2 encode combinations of orientations. *Nature Neuroscience*, 10(10):1313–1321, 2007. doi: 10.1038/nn1975. URL https://doi.org/10.1038/nn1975.

Irving Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987. doi: 10.1037/0033-295X.94.2.115.

Charles Cadieu, Minjoon Kouh, Anitha Pasupathy, Charles E. Connor, Maximilian Riesenhuber, and Tomaso Poggio. A model of V4 shape selectivity and invariance. *Journal of Neurophysiology*, 98(3):1733–1750, 2007. doi: 10.1152/jn.01265.2006.

Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib Majaj, and James J. DiCarlo. Deep neural networks rival the representation of primate it cortical neurons. *Proceedings of the National Academy of Sciences*, 111(23):8519–8524, 2014.

T. Caelli and B. Julesz. On perceptual analyzers underlying visual texture discrimination: Part i. *Biological Cybernetics*, 28(3):167–175, 1978. doi: 10.1007/BF00337138.

Riccardo Caramellino, Eugenio Piasini, Andrea Buccellato, Anna Carboncino, Vijay Balasubramanian, and Davide Zoccolan. Rat sensitivity to multipoint statistics is predicted by efficient coding of natural scenes. *eLife*, 10:e72081, 2021. doi: 10.7554/eLife.72081.

Matteo Carandini, Jonathan B. Demb, Valerio Mante, David J. Tolhurst, Yang Dan, Bruno A. Olshausen, Jack L. Gallant, and Nicole C. Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–10597, 2005. doi: 10.1523/JNEUROSCI.3726-05.2005.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. doi: 10.1109/ICCV48922.2021.00951.

Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Stephen V. David, Benjamin Y. Hayden, and Jack L. Gallant. Spectral receptive field properties explain shape selectivity in area V4. *Journal of Neurophysiology*, 96(6): 3492–3505, 2006. doi: 10.1152/jn.00575.2006.

Ludovica de Paolis, Fabio Anselmi, Alessio Ansuini, and Eugenio Piasini. Perceptual misalignment of texture representations in convolutional neural networks. *arXiv preprint arXiv:2604.01341*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

Robert Desimone and Susan J. Schein. Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. *Journal of Neurophysiology*, 57(3):835–868, 1987. doi: 10.1152/jn.1987.57.3.835.

James J. DiCarlo and David D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, 2007. doi: 10.1016/j.tics.2007.06.010.

James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. doi: 10.1016/j.neuron.2012.01.010.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Christina Enroth-Cugell and J. G. Robson. The contrast sensitivity of retinal ganglion cells of the cat. *The Journal of Physiology*, 187(3):517–552, 1966. doi: 10.1113/jphysiol.1966.sp008107.

Jeremy Freeman and Eero P. Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 14(9):1195–1201, 2011. doi: 10.1038/nn.2889.

Jeremy Freeman, Corey M. Ziemba, David J. Heeger, Eero P. Simoncelli, and J. Anthony Movshon. A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, 16(7):974–981, 2013.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36 (4):193–202, 1980. doi: 10.1007/BF00344251.

R. Gattass, C. G. Gross, and J. H. Sandell. Visual topography of V2 in the macaque. *The Journal of Comparative Neurology*, 201(4):519–539, 1981. doi: 10.1002/cne.902010405.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. doi: 10.1523/JNEUROSCI.5023-14.2015.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, 2023. doi: 10.1109/TPAMI.2022.3152247.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Jay Hegdé and David C. Van Essen. Selectivity for complex shapes in primate visual area V2. *Journal of Neuroscience*, 20(5):RC61, 2000. doi: 10.1523/JNEUROSCI.20-05-j0001.2000. URL https://doi.org/10.1523/JNEUROSCI.20-05-j0001.2000.

Ann M. Hermundstad, John J. Briguglio, Mary M. Conte, Jonathan D. Victor, Vijay Balasubramanian, and Gašper Tkačik. Variance predicts salience in central sensory processing. *eLife*, 3:e03722, 2014. doi: 10.7554/eLife.03722.

David H. Hubel and Torsten N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591, 1959. doi: 10.1113/jphysiol.1959.sp006308.

David H. Hubel and Torsten N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962. doi: 10.1113/jphysiol.1962.sp006837.

David H. Hubel and Torsten N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, 1968. doi: 10.1113/jphysiol.1968.sp008455.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456. PMLR, 2015.

Minami Ito and Hidehiko Komatsu. Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *Journal of Neuroscience*, 24(13):3313–3324, 2004. doi: 10.1523/JNEUROSCI.4364-03.2004. URL https://doi.org/10.1523/JNEUROSCI.4364-03.2004.

Bela Julesz. Visual pattern discrimination. *IRE Transactions on Information Theory*, 8:84–92, 1962.

Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981. doi: 10.1038/290091a0.

Bela Julesz, E. N. Gilbert, L. A. Shepp, and H. L. Frisch. Inability of humans to discriminate between visual textures that agree in second-order statistics revisited. *Perception*, 2(4):391–405, 1973.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PNAS*, 111(23):8619–8624, 2014.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s):1–41, 2022. doi: 10.1145/3505244.

Nikolaus Kriegeskorte. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446, 2015. doi: 10.1146/annurev-vision-082114-035447.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc., 2012.

Hendrik G. J. M. Kuypers, Maria K. Szwarcbart, Mortimer Mishkin, and Haldor E. Rosvold. Occipitotemporal corticocortical connections in the rhesus monkey. *Experimental Neurology*, 11:245–262, 1965. doi: 10.1016/0014-4886(65)90016-3.

Grace W. Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, 33(10):2017–2031, 2021. doi: 10.1162/jocn_a_01544.

Nikos K. Logothetis and David L. Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19:577–621, 1996. doi: 10.1146/annurev.ne.19.030196.003045.

Jitendra Malik and Pietro Perona. Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A*, 7(5):923–932, 1990. doi: 10.1364/JOSAA.7.000923.

Bartlett W. Mel. Seemore: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, 9(4): 777–804, 1997. doi: 10.1162/neco.1997.9.4.777.

J. A. Movshon, I. D. Thompson, and D. J. Tolhurst. Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *The Journal of Physiology*, 283:53–77, 1978a. doi: 10.1113/jphysiol.1978.sp012488.

J. A. Movshon, I. D. Thompson, and D. J. Tolhurst. Receptive field organization of complex cells in the cat's striate cortex. *The Journal of Physiology*, 283:79–99, 1978b. doi: 10.1113/jphysiol.1978.sp012489.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193. arXiv:2304.07193.

Anitha Pasupathy and Charles E. Connor. Responses to contour features in macaque area V4. *Journal of Neurophysiology*, 82(5):2490–2502, 1999. doi: 10.1152/jn.1999.82.5.2490.

Anitha Pasupathy and Charles E. Connor. Shape representation in area V4: position-specific tuning for boundary conformation. *Journal of Neurophysiology*, 86(5):2505–2519, 2001. doi: 10.1152/jn.2001.86.5.2505.

Anitha Pasupathy, Dina V. Popovkina, and Taekjun Kim. Visual functions of primate area V4. *Annual Review of Vision Science*, 6:363–385, 2020. doi: 10.1146/annurev-vision-030320-041306.

Karalyn Patterson, Peter J. Nestor, and Timothy T. Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8:976–987, 2007. doi: 10.1038/nrn2277.

Eugenio Piasini. Metex - maximum entropy textures, v1.2.0, 2021. URL https://zenodo.org/records/18500242.

Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000a. doi: 10.1023/A:1026553619983.

Javier Portilla and Eero P. Simoncelli. textureSynth: Matlab code for analyzing and synthesizing visual textures, 2000b. https://github.com/LabForComputationalVision/textureSynth.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999. doi: 10.1038/14819.

Maximilian Riesenhuber and Tomaso Poggio. Models of object recognition. *Nature Neuroscience*, 3(Suppl):1199–1204, 2000. doi: 10.1038/81479.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Nicole C. Rust and James J. DiCarlo. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *Journal of Neuroscience*, 30(39):12978–12995, 2010. doi: 10.1523/JNEUROSCI.0179-10.2010.

Anita M. Schmid, Keith P. Purpura, Ifije E. Ohiorhenuan, Ferenc Mechler, and Jonathan D. Victor. Subpopulations of neurons in visual area V2 perform differentiation and integration operations in space and time. *Frontiers in Systems Neuroscience*, 3:15, 2009. doi: 10.3389/neuro.06.015.2009. URL https://doi.org/10.3389/neuro.06.015.2009.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018. URL https://www.biorxiv.org/content/10.1101/407007v2.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pages 1–14, 2015.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL https://www.jmlr.org/papers/v15/srivastava14a.html.

C. Szegedy, W. Liu, y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

Keiji Tanaka. Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–139, 1996. doi: 10.1146/annurev.ne.19.030196.000545.

Tiberiu Tesileanu, Mary M. Conte, John J. Briguglio, Ann M. Hermundstad, Jonathan D. Victor, and Vijay Balasubramanian. Efficient coding of natural scene statistics predicts discrimination thresholds for grayscale textures. *eLife*, 9:e54347, 2020. doi: 10.7554/eLife.54347.

Tiberiu Tesileanu, Eugenio Piasini, and Vijay Balasubramanian. Efficient processing of natural scenes in visual cortex. *Frontiers in Cellular Neuroscience*, 16:1006703, 2022. ISSN 1662-5102. doi: 10.3389/fncel.2022.1006703.

Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996. doi: 10.1038/381520a0.

TorchVision maintainers and contributors. TorchVision: PyTorch's Computer Vision library, November 2016. URL https://github.com/pytorch/vision.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

J. D. Victor and M. M. Conte. Local image statistics: maximum entropy constructions and perceptual salience. *Optical Society of America*, 29:1313–1345, 2012.

Jonathan D. Victor, Mary M. Conte, and Charles F. Chubb. Textures as probes of visual processing. *Annual Review of Vision Science*, 3:275–296, 2017.

Thomas S. A. Wallis, Christima M. Funke, Alexander S. Ecker, Leon A. Gatys, Felix A. Wichmann, and Mathias Bethge. A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of Vision*, 17: 1–29, 2017.

Ahmad Waseem, Pietro Ruiu, Seth Nixon, Andrea Lagorio, and Massimo Tistarelli. Self-attention as the backbone: A survey on vision transformers. *Computer Vision and Image Understanding*, 269:104791, 2026. doi: 10.1016/j.cviu.2026.104791.

Ben D. B. Willmore, Ryan J. Prenger, and Jack L. Gallant. Neural representation of natural images in visual area V2. *Journal of Neuroscience*, 30(6):2102–2114, 2010. doi: 10.1523/JNEUROSCI.4099-09.2010. URL https://doi.org/10.1523/JNEUROSCI.4099-09.2010.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23): 8619–8624, 2014.

Mirko Zanon, Bastien S. Lemaire, Eugenio Piasini, Riccardo Caramellino, Caroline Nallet, Vijay Balasubramanian, Judit Gervain, Davide Zoccolan, and Giorgio Vallortigara. Predisposed and learned preferences for multipoint visual statistics in visually naïve newly hatched chicks. *bioRxiv*, 2025. bioRxiv preprint.

Corey M. Ziemba, Jeremy Freeman, J. Anthony Movshon, and Eero P. Simoncelli. Selectivity and tolerance for visual texture in macaque v2. *Proceedings of the National Academy of Sciences*, 113(22), 2016.

CHAPTER 2

A Variational Autoencoder for texture synthesis

In this chapter I describe how I developed and curated the full pipeline of a deep learning model for texture synthesis, from literature review and dataset preparation to model design, training, and evaluation. The model combines a Variational Autoencoder (VAE) with Gatys' Gram matrix representations with the goal of generating naturalistic textures in a structured latent space. Although the full approach did not succeed, the experiments isolated the main limitations of the texture representation. At the same time, I successfully reproduced and validated the Gatys algorithm. This result led to a more successful follow-up study focused directly on CNN based texture representations.

Statement of contribution: I developed the framework, conducted experiments, analyzed the results and wrote this chapter. Paolo Muratore provided technical assistance for the development of the framework. Eugenio Piasini supervised the work.

2.1 Abstract

According to the efficient coding principle, the mammalian visual system is adapted to the statistical structure of natural visual scenes. Visual textures provide a useful experimental framework to investigate this hypothesis, but existing psychophysical studies have often relied on artificial stimuli defined by low-order correlations. Here, I explored a new approach to texture synthesis aimed at combining the naturalistic representations of convolutional neural networks with the structured latent space of a Variational Autoencoder. The proposed model couples a convolutional VAE with a pre-trained VGG network and uses Gram matrices of deep feature activations, following Gatys et al. (2015), to define texture similarity during training.

I first validated the individual components of the framework. Our implementation of the Gatys algorithm successfully reproduced naturalistic textures, and the autoencoding backbone was able to reconstruct simple benchmarking images and, in its deterministic form, images from the Describable Textures Dataset. However, the complete VAE-based architecture did not generate high-quality naturalistic textures when trained on natural scenes or texture datasets. In particular, reconstruction quality deteriorated substantially when the variational regularization was introduced, suggesting that compressing complex deep texture statistics into a simple Gaussian-regularized latent space imposes a strong limitation on the model.

Although the original objective of constructing a controllable latent manifold of naturalistic textures was therefore not achieved, the experiments provided a useful negative result. By progressively isolating the different components of the model, I identified the Gram-matrix representation itself as the most robust part of the framework. This motivated the subsequent investigation of the Gatys representation directly, leading to a more successful line of work on the statistical structure of texture representations in deep neural networks.

2.2 Introduction

According to the efficient coding principle, a sensory system conforms to the statistics of its natural input (Barlow, 2001; Simoncelli, 2003). In the case of the mammalian visual system, efficient coding predicts that it conforms to the statistics of natural visual scenes, the natural environment in which mammals, including human, shape their visual experience across their life and across evolution. Efficient coding, therefore, entails that neuronal encoding in the visual systems adapts strategically to the statistics of visual stimuli, retaining relevant information and discarding the rest (Hermundstad et al., 2014; Sterling and Laughlin, 2017; Tesileanu et al., 2022). In practice, the visual system is expected to attribute less importance to redundant features, and to respond more strongly to unexpected visual stimuli.

However, Hermundstad et al. (2014) put forward a "variance is salience" hypothesis for central visual processing. This hypothesis argues that the neural processing in the intermediate areas of the visual cortex, which is responsible for texture processing, should respond to high-variance stimulus features and should receive a higher proportion of resources. The authors therefore claim that stimuli characterized by high variance should possess higher perceptual salience. The "variance is salience" hypothesis stands in contrast with classic efficient coding predictions, that downweight high-variance stimulus features in peripheral processing (a direct application of the theory was first described by van Hateren (1992)). Hermundstad et al. (2014) test this hypothesis by using binary synthetic texture images generated with the model by Victor and Conte (2012). Later, Tesileanu et al. (2020) extend this efficient coding account of texture perception from binary textures to grayscale images. They test the "variance is salience" hypothesis with a psychophysics texture segregation task on human observers and show evidence that sensitivity to grayscale texture statistics is adapted to natural scene variability, confirming the results obtained by Hermundstad et al. (2014). The results obtained with human subjects are consistent with those of other studies with rodents (Caramellino et al., 2021) and chicks (Zanon et al., 2025).

Visual textures have been defined as "a family of visual patterns that share certain local regularities" (Julesz, 1981), and as seen in the studies mentioned above, they represent a tool to understand how efficient coding applies to the mammalian visual system. In fact, textures distill the statistical properties of natural images, as they satisfy both local and global statistical constraints (Tkačik et al., 2010). Julesz was the first one to propose not only a formal definition of textures, but also to hypothesize that textures can be statistically determined. His work inspired the development of several texture synthesis models because, as suggested in Julesz (1962), synthetic textures are a better method for obtaining textures stimuli compared to natural images crops, because contextual information in natural image is a source of bias.

The experimental studies discussed above rely on texture synthesis, which is a group of mathematical methods to generate artificial visual textures that were inspired by the seminal studies of Julesz, and in particular to his conjecture according to which textures that share up to second-order statistics can't be distinguished (Julesz et al., 1973). Scientific literature reports several kinds of models that applied the conjecture by defining texture representations as correlations among low-order statistics. These models include parametric methods, (Portilla and Simoncelli, 2000; Victor and Conte, 2012), convolutional neural networks (CNNs) (Gatys et al., 2015; Heitz et al., 2021; Sendik and Cohen-Or, 2017; Snelgrove, 2017; Ulyanov et al., 2017; Vacher et al., 2020), generative adversarial networks (GANs) (Jetchev et al., 2016), and attention-based networks (Guo et al., 2022).

The performance of texture synthesis models has been evaluated by humans through several behavioral experiments (Portilla and Simoncelli, 2000; Victor and Conte, 2012; Wallis et al., 2017), as well as neuroimaging studies (Henderson et al., 2023; Jagadeesh and Gardner, 2021). Three of these models are particularly relevant for motivating the present study: Portilla and Simoncelli (2000) developed a parametric model that relies on the parameters of wavelet coefficients, and synthetic textures are generated to statistically match the original ones (Portilla and Simoncelli, 2000); Victor and Conte (2012) developed a method that samples maximum-entropy textures with fixed correlation statistics among local neighborhoods; Gatys et al. (2015) used a CNN to compute correlations between feature maps and generated textures with gradient descent on the pixel space to reproduce the correlations in the input images. Taken together, these three models are limited to generating textures that appear coarse grained and unnatural, or to operate in spaces whose parameters are difficult to interpret: on the one hand the model by Victor and Conte (2012) provides an elegant parameterization of texture space suitable for highly-controlled experiments, but produces textures that are perceptually distant from those found in natural images; on the other hand, Gatys et al. (2015) model generates naturalistic-looking textures but lacks explanations of the model's internal representation and provides fewer affordances towards investigating the perceptual properties of textures. As a consequence, results in testing efficient coding have been limited to artificial-looking visual textures characterized by low-order correlations among neighboring pixels (Caramellino et al., 2021; Hermundstad et al., 2014; Tesileanu et al., 2022; Victor and Conte, 2012). The model of Portilla and Simoncelli (2000), while accounting for both pixel-level local correlations and for global correlations based on the spectral analysis of the input image, can't fully overcome the challenges posed by the other two models. In fact, such framework only partially preserves the local structure of the original image as long as it displays a regular and periodic pattern. This results in grayscale textures that do not look naturalistic (Freeman et al., 2013; Portilla and Simoncelli, 2000; Wallis et al., 2017).

Overall, existing methods for texture synthesis limit out ability to test efficient coding, either because they rely on high dimensional parameterizations that are hard to interpret and manipulate, or because they produce overly simplistic stimuli that in practice do not look like naturalistic textures.

Our goal is to develop a new method that would overcome such limitations by generating naturalistic-looking textures characterized by nonlinear multi-scale representations as in Gatys et al. (2015), while relying on a mathematically tractable latent space as in Victor and Conte (2012). I propose a new model which combines a Variational Autoencoder (VAE) (Doersch, 2016; Kingma and Welling, 2014) and a pre-trained CNN (VGG-19 (Simonyan and Zisserman, 2015)). VAEs are a type of generative model, along with Generative Adversarial Networks (GAN) (Goodfellow, 2017; Goodfellow et al., 2014) and latent diffusion models (Rombach et al., 2022). These models are widely used for many computer vision tasks, such as image generation (Yu et al., 2022), image reconstruction (Takagi and Nishimoto, 2023), scene prediction Walker et al. (2016), representation learning (Bouchacourt et al., 2018; van den Oord et al., 2017), neural texture synthesis (Bergmann et al., 2017), and image-caption learning (Pu et al., 2016).

2.3 Methods

2.3.1 Model architecture

Our model architecture backbone consists of a convolutional VAE, which I use as image generator, and by VGG-19 weights pretrained on object classification which I use to obtain the texture representations to be imposed on the generated image. In practice, our model is designed to generate images that show new, synthetic examples of the visual textures found in the input images. The VAE consists of a probabilistic encoder that approximates the posterior distribution,

$$\mathcal{E} = q_\phi(z|x), \tag{2.1}$$

and a probabilistic decoder that parametrizes the likelihood,

$$\mathcal{D} = p_\theta(x|z). \tag{2.2}$$

The encoder $\mathcal{E}$ accepts as an input images $x \in P(X)$ of dimensionality $x \in \mathcal{R}^{H\times W\times C} = R^{224\times 224\times 3}$, and converts them into the so-called latent code $z \in P(Z)$, which is a vector representation of the features of the image features. The decoder $\mathcal{D}$ generates new image $\hat{x} \in \mathcal{R}^{224\times 224\times 3}$ from $P(Z)$ by sampling a Gaussian distribution $\mathcal{N} \sim (0, 1)$ via reparametrization, defined as:

$$z = \mu + \sigma \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{2.3}$$

where ϵ is the stochastic term sampled from a Gaussian.

The VAE is trained to minimize the distance between the encoder probability distribution (see Equation 2.1) and the Gaussian distribution.

At each generative iteration the original and the generated images are given to a pretrained VGG-19 to obtain texture features based on the technique in Gatys et al. (2015), namely computing the Gram Matrices of VGG-19 feature maps for each original-generated image pairs. The Gram Matrices are computed as the dot product of the set of feature maps in each convolutional layer:

$$G^l_{ij} = \sum_{m=1}^{M_l} F^l_{im} F^l_{jm} \tag{2.4}$$

where F^l_{im} is the is the activation of feature map i at spatial position m in layer l, discarding the spatial dimensions and only retaining the statistical features of the textures detected by the filters. As in Gatys et al. (2015), I obtain 5 original-reconstructed Gram Matrix pairs for 5 convolutional layers at different depths.

2.3.2 Loss

A standard VAE image generator optimizes the following loss:

$$\mathcal{L}_{total} = -\sum_{i=1}^{N} E_{z_i\sim q_\phi(z_i|x_i)}[\frac{1}{M}\sum_{m=1}^{M}(x_{im} - \hat{x}_{im})^2] - \beta\,\mathrm{KL}(q_\phi(z_i|x_i)||p(z_i)) \tag{2.5}$$

where the last term is the Kullback-Leibler (KL) divergence between the latent code distribution $q_\phi(z_i|x_i)$ and the Gaussian distribution for which $\mathcal{E}$ was reparameterized (see Equation 2.3); the first term is the Mean Squared Error (MSE) in the pixel space of the

input x_i and of the generated image $\hat{x}_i$, taken in expectation over $q_\phi(z_i|x_i)$ (in practice, this is estimated using a single sample from $q_\phi(z|x)$). In principle, the MSE loss computed on the pixel space pushes the model to reconstruct images as identical as possible to the original images. In this way, the model would not actually learn to represent the distribution of the textures in the input data. For this reason, in the total loss I substitute pixel MSE loss term with a MSE loss computed between pairs of Gram Matrices to define the perceptual difference between original and generated textures:

$$\mathcal{L}_{gram} = \sum_{n=1}^{L} \frac{1}{K_{l_n}^2} \sum_{k,j=1}^{K_{l_n}} (\hat{G}_{kj}^{l_n} - G_{kj}^{l_n})^2 \tag{2.6}$$

where $\{l_1, \ldots, l_L\}$ is the set of network layers used to compute the loss, and K_l is the number of feature maps in layer l.

Therefore the total loss of our model becomes:

$$\mathcal{L}_{total} = -\sum_{i=1}^{N} E_{z_i \sim q_{\phi(z_i|x_i)}} [\sum_{n=1}^{L} \frac{1}{K_{l_n}^2} \sum_{k,j=1}^{K_{l_n}} (\hat{G}_{kj}^{(i)l_n} - G_{kj}^{(i)l_n})^2] - \beta \, \mathrm{KL}(q_\phi(z_i|x_i)||p(z_i)) \tag{2.7}$$

where the KL divergence acts as a regularizer term for model training, and the MSE term guides the generation of images by minimizing the difference between the Gram matrices of the original image and of the reconstructed image. A visual representation of the model and of the specifics of the parametrization are available in Appendix Figure 2.5, Tables 2.1, 2.2.

At each training iteration, an input image is passed to the VAE encoder, which computes the latent code, and to VGG-19 to extract the Gram matrices as texture representations. From the latent code, the decoder samples a synthetic image which also gets passed to VGG-19 for extraction of the Gram representation. The two Gram representations are used to compute the loss 2.7, which is minimized using Adam. The purpose of this process is to lead the model to generate new image exemplars that display textures characterized by the same Gram representation as the original image.

In the training process the model relies on two main hyperparameters that I determine with a cross-validation analysis: the dimensionality of the latent code, and the weight assigned to the KL loss (β (see Equation 2.7)), which balance the MSE and the KL loss.

2.4 Experiments

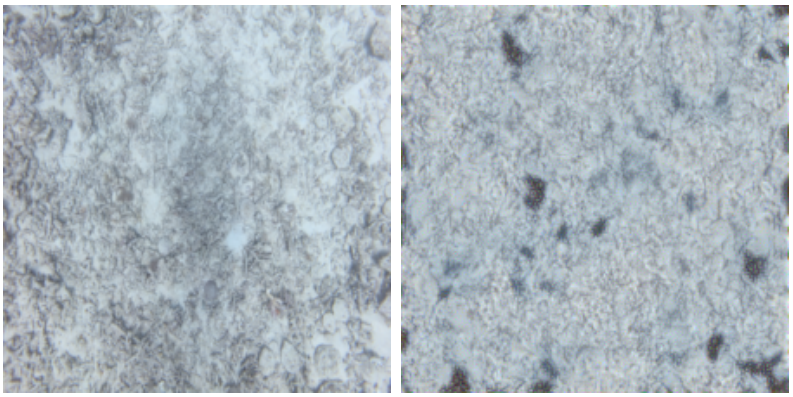
I started by implementing the Gatys et al. (2015) algorithm for texture synthesis and testing its quality and result reproducibility, as a fundamental step to ensure the extraction of texture representations for our model. Subsequently, I trained the full model using multiple datasets, including two benchmarking datasets, one custom dataset, a dataset of objects, and a dataset of naturalistic textures. Below I report the pipelines that I followed for the training process.

2.4.1 Evaluation of the Gatys algorithm

In figure 2.1 I show reconstructed textures from original pictures at 27.000 optimization steps with Adam. This result shows that our implementation of the Gatys algorithm produces high quality textures that compare to those generated in the original work (see Gatys et al. (2015), Figure 2). I therefore proceeded to train our model with this implementation.



(a) Original photo of boats vs. reconstruction.



(b) Original photo of a marine landscape vs. reconstruction.

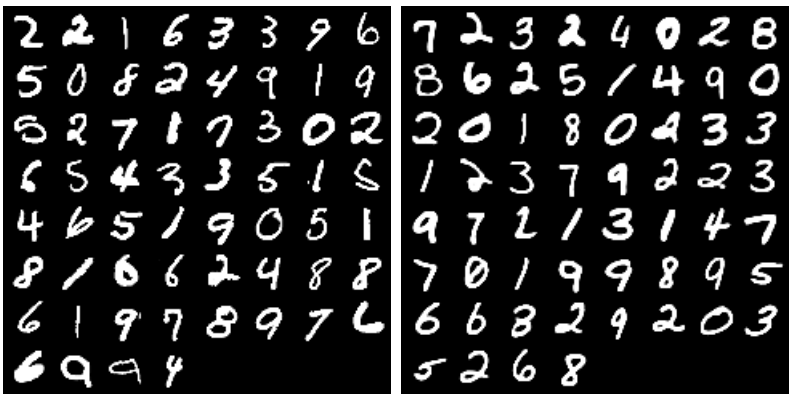
Figure 2.1: **Replication of the Gatys algorithm** with two original images (first column) and corresponding reconstructions (second column).

2.4.2 Benchmarking datasets

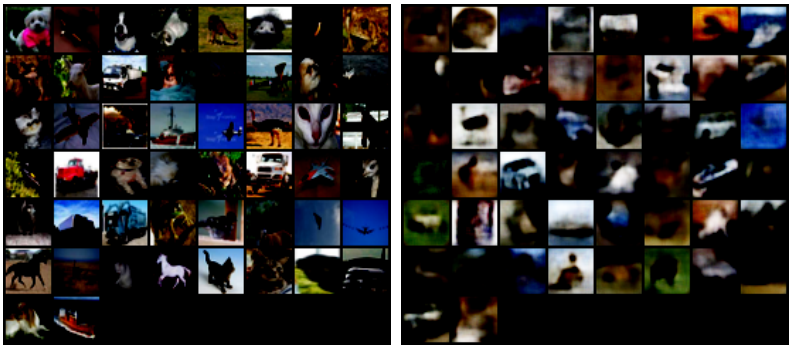
I first evaluated the performance of the model with MNIST (LeCun et al.) and CIFAR-10 (Krizhevsky, 2009). MNIST is a large supervised database of handwritten digits, divided in 10 classes. It features 60.000 binary images of numbers of dimensionality 28×28 pixels. CIFAR-10 is a large supervised dataset of colored images, divided in 10 semantic classes of objects, for a total of 60.000 images. The dimensionality of each image is 32×32 pixels. Both MNIST and CIFAR-10 are frequently used as benchmarking databases to evaluate the performance of machine learning models for vision or computer vision systems. In fact, MNIST features images with 1 channel and no semantic content, while CIFAR-10 contains images of objects in RGB with low spatial resolution, making them very convenient to test the efficacy of the model in early training experiments.

I used both MNIST and CIFAR-10 to train the VAE backbone with the pixel loss (Equation 2.5), setting $\beta = 1$. I furthermore set the dimensionality of the latent code $Z = 10$, to match the number of classes in both datasets. A summary of the model configuration is reported in Supplementary Table 2.3.

In Figure 2.2 I report the result of our first training experiment, observing the generative capability of the VAE with this specific configuration. In the case of MNIST, the model reaches visual quality comparable to the original images, while in the case of CIFAR-10 the quality is lower, although shape and color are largely preserved. I decided to proceed by testing the Gram loss in the other datasets.



(a) MNIST original vs. reconstruction.



(b) CIFAR-10 original vs. reconstruction.

Figure 2.2: **MNIST and CIFAR-10 training experiment.** Random samples of original images (first column) and reconstructions (second column). This figure is for illustrative purposes to examine the quality of the output with respect to the quality of the input.

2.4.3 Large unlabeled datasets

I proceeded by implementing the full training loss in the space of Gram texture representations in Equation 2.7. I first used a custom dataset of unpublished naturalistic landscapes (courtesy of Sam Reiter, OIST), and successively I used the Natural Scenes Dataset (NSD) (Allen et al., 2022). The custom dataset features ~ 7000 images of naturalistic pictures of underwater landscapes of very high visual resolution, around 5000×2000) pixels. I therefore resized and cropped all the images to 224×224 and applied a number of transformations to the images in order to both fit the configuration of the model (see Supplementary Table 2.4), and as a data augmentation step to maximize the number of novel training samples. Both datasets are unlabeled but are characterized by naturalistic images. NSD (Allen et al., 2022) is a large dataset of 60.000 images of 224×224 pixels, featuring everyday and naturalistic scenes. In addition to the image stimuli of our interest, NSD features neural data collected with different methods of human observers exposed to the image.

I chose these datasets because I aimed to train a generative model to synthesize textures, and therefore were interested in images that do not feature exclusively objects (as opposed for instance to popular datasets often used for object classification, such as ImageNet, Deng et al. (2009); Russakovsky et al. (2015)). For this reason I wanted to avoid any semantic categorization to influence the formation of a latent distribution, as well as object image features to influence the learning process. Naturalistic landscapes or scenes

could be a useful type of data to achieve this goal.

I first assessed the texture generation ability of the model by training it with the custom dataset and the setup reported in Supplementary Table 2.4. An exemplar of input image and of reconstruction performed with full loss (see Eq. 2.7) is shown in Figure 2.3a. The reconstruction features groups of artifacts that do not reflect the type of result that I was aiming at: high quality naturalistic-looking textures as in Gatys et al. (2015). I attributed this result to the scarcity of input data, as well as their low variability (many images from out custom dataset are similar to one another). I therefore decided to proceed with training on NSD, and a new model setup (see Supplementary Table 2.5). I aimed at increasing the quantity and the inter-variability of the training data. I furthermore implemented a learning rate decay scheduler, and I added a weight α to the MSE on the gram representation of the loss, which becomes:

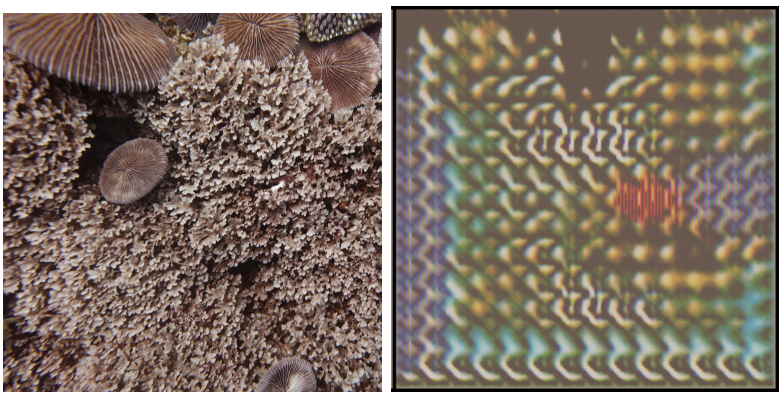
$$\begin{aligned}\mathcal{L}_{total} = -\sum_{i=1}^{N} E_{z_i \sim q_{\phi(z_i|x_i)}} \big[(1-\alpha)[\frac{1}{M^2}\sum_{m=1}^{M}(x_m - \hat{x_m})^2] + \\ + \alpha[\sum_{n=1}^{L}\frac{1}{K_{l_n}^2}\sum_{k,j=1}^{K_{l_n}}(\hat{G}_{kj}^{(i)l_n} - G_{kj}^{(i)l_n})^2]\big] - \beta\,\mathrm{KL}(q_\phi(z_i|x_i)||p(z_i))\end{aligned} \tag{2.8}$$

where α is initially set to 0, and subject to a scheduler. For 4 training days it is kept stable at 0 such that the pixel loss takes over the reconstruction to guide the model towards learning the space of pixels and generating naturalistic images; afterwards, α is interpolated to smoothly transition to 1, to gradually bring the model to learn the gram representation of the images and actually reconstruct naturalistic textures. I show the result of this experiment in Figure 2.3b. I observe that, although the pixel level reconstruction improved to the point of showing object-like shapes and colors, the images are still overall of poor quality and present artifacts that do not resemble texture representations. By inspecting the loss trend during training, I see that this result emerges when the losses value have decreased significantly (see Supplementary Figures 2.6). Due to the training difficulties encountered with the datasets explored so far, I then decided to focus on a texture image dataset, which I expected to be more naturally suited to our goals.

2.4.4 Texture dataset

This training experiment was conducted with the model configuration summarized in Supplementary table 2.6 and with the Describable Texture Dataset (DTD Cimpoi et al. (2014)). DTD is a database that collects 5640 images of naturalistic textures divided into 47 labels. The labels are assigned with an independent linguistic task based on human judgment agreement, and are based on an independent study by Bhushan et al. (1997). This dataset is useful for training our model because it explicitly encodes images of textures which are also divided in semantic classes, so it contains texture information both at the perceptual level and at the class level. Therefore, it could potentially support generation of novel textures as well as the learning of a disentangled latent space.

To address the training issues highlighted in the previous session, I simplified the architecture to the bare minimum – a (non-variational) autoencoder (AE) in pixel space – and then to gradually add components to isolate the emergence of the issue. Since DTD contains a relatively small number of images, I also implemented several data augmentation operations.



(a) Custom dataset original vs. reconstruction.



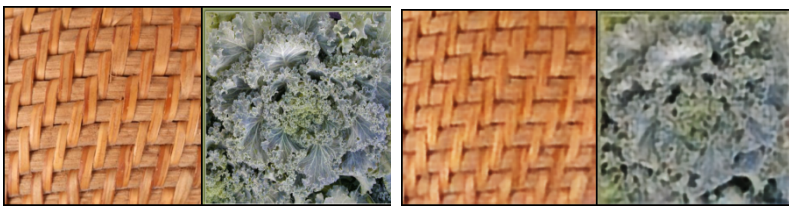
(b) NSD original vs. reconstruction.

Figure 2.3: **Custom dataset and NSD training experiment.** Random samples of original images (first column) and reconstructions (second column). This figure is for illustrative purposes to examine the quality of the output with respect to the quality of the input.

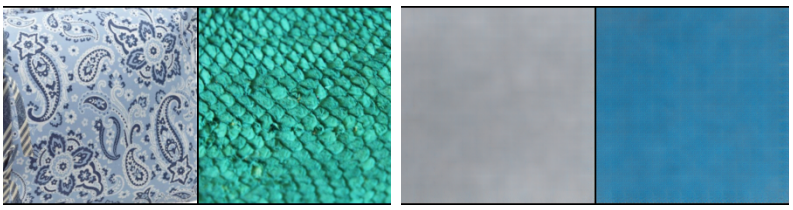
Figure 2.4a shows two images from DTD reconstructed pixel by pixel by the AE, the simplest version of our VAE backbone, that optimizes the following loss:

$$\mathcal{L}_{pixel} = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 \tag{2.9}$$

The reconstruction is satisfactory as it preserves good image quality and semantic content, even if if the details are not as perceptually good as the original pictures. I therefore proceeded to add the KL regularization loss, effectively switching to a VAE backbone, to test the generative model by keeping β very low and Z as the same number of DTD classes (see Table 2.6 for implementation details). I kept the pixel loss as in Eq 2.9. The results in Figure 2.4b shows that at this level of complexity the models shows inability to generate novel images by sampling the input dataset. DTD is much more limited with respect to NSD, and that it does not have enough internal variability due to the fact that only textures are represented. In addition, an inspection of the loss trend (see Figure 2.7) reveals that VAE learning is very unstable. I therefore concluded that the generative ability of the model in this experiment is affected by the lack of samples in the input data, the dimensionality of Z, which is not optimized to capture a tractable representations of the textures in the dataset, and by the instability of the architecture.



(a) AE DTD original vs. reconstruction.



(b) VAE DTD original vs. reconstruction.

Figure 2.4: **DTD training experiments.** Original images (first column) and corresponding reconstructions (second column). In both the AE and the VAE versions of the model I trained for 100 epochs and kept the parameter configurations displayed in Table 2.1.

2.5 Discussion and conclusions

In this chapter, I investigated whether the statistical texture representation introduced by Gatys et al. (2015) could be combined with the structured latent space of a VAE to obtain a generative model of naturalistic textures. The motivation was to combine two desirable properties that are usually found separately in existing texture synthesis approaches: the ability of Gram matrix-based methods to synthesize visually rich and naturalistic textures, and the explicit low-dimensional latent representation provided by VAEs. Such a representation could, in principle, provide a continuous texture space whose geometry could be explored and ultimately compared with perceptual representations measured in human observers.

As a first validation step, I implemented the texture synthesis algorithm of Gatys et al. (2015) independently from the VAE. The reconstructed textures showed that our implementation of the Gram matrix objective was capable of producing high-quality synthetic images, qualitatively reproducing the behavior of the original method. I then assessed the reconstruction capabilities of the VAE backbone using MNIST and CIFAR-10. While the network successfully reconstructed the simpler MNIST images and retained some coarse visual information for CIFAR-10, reconstruction quality decreased substantially as the complexity of the input images increased.

I subsequently combined the two components and trained the complete model on progressively more appropriate datasets, including a custom dataset of natural scenes, the NSD, and the DTD. Across these experiments, however, the proposed model did not reach the intended behavior. In particular, optimizing the VAE jointly through pixel space reconstruction, latent-space regularization, and Gram matrix matching did not result in high quality naturalistic textures comparable to those obtained when the Gatys optimization was applied directly in image space. Increasing the amount and variability of the training data, modifying the relative contribution of the different loss terms, and introducing a gradual transition from pixel reconstruction to Gram matrix optimization improved some aspects of the reconstructions but did not solve the fundamental limitation.

These results suggest that a straightforward combination of a VAE with a Gram matrix reconstruction objective is not sufficient to construct the desired latent texture manifold. The two components impose substantially different constraints on the generated images. The VAE must compress each image through a low-dimensional probabilistic bottleneck while simultaneously maintaining a latent distribution close to the chosen prior, whereas the Gram matrix objective constrains high dimensional statistics extracted at several depths of a pretrained CNN. Although many different images may share similar Gram representations, there is no guarantee that these equivalence classes can be represented accurately by the comparatively restricted latent space learned by the VAE. The optimization therefore needs to satisfy simultaneously image reconstruction, latent regularization, and preservation of multiscale texture statistics, and our experiments indicate that the architecture and objective considered here were not sufficient to achieve this balance.

Importantly, the failure of the complete model does not invalidate the use of deep convolutional representations for studying texture statistics. On the contrary, one of the clearest outcomes of this project was the successful independent implementation and validation of the Gatys texture synthesis framework. The contrast between the reliable behavior of direct Gram matrix optimization and the poorer performance obtained after embedding the same objective within a VAE suggested that the Gram representation itself was a more promising object of study than the particular latent-variable architecture explored here.

This observation motivated the subsequent studies developed from this project, in which I focused directly on the properties of the Gatys representation of texture statistics. Removing the additional constraints imposed by the VAE made it possible to investigate the Gram representation in a more controlled setting and ultimately led to successful results. Thus, although the original objective of constructing a VAE based latent manifold of naturalistic textures was not achieved, the negative result played an important methodological role: it identified which component of the proposed framework was robust, clarified the limitations of the combined architecture, and redirected the project toward a more productive analysis of Gram-based texture representations.

2.6 Appendix

2.6.1 Model architecture

VAE	
Input image ($224 \times 224 \times 3$)	Latent code
conv-64 conv-64	deconv-64 deconv-64
maxpool	
conv-128 conv-128	deconv-128 deconv-128
maxpool	
conv-256	deconv-256
maxpool	
Latent code	Output image $224 \times 224 \times 3$

Table 2.1: **Architecture of the Variational Autoencoder backbone.**

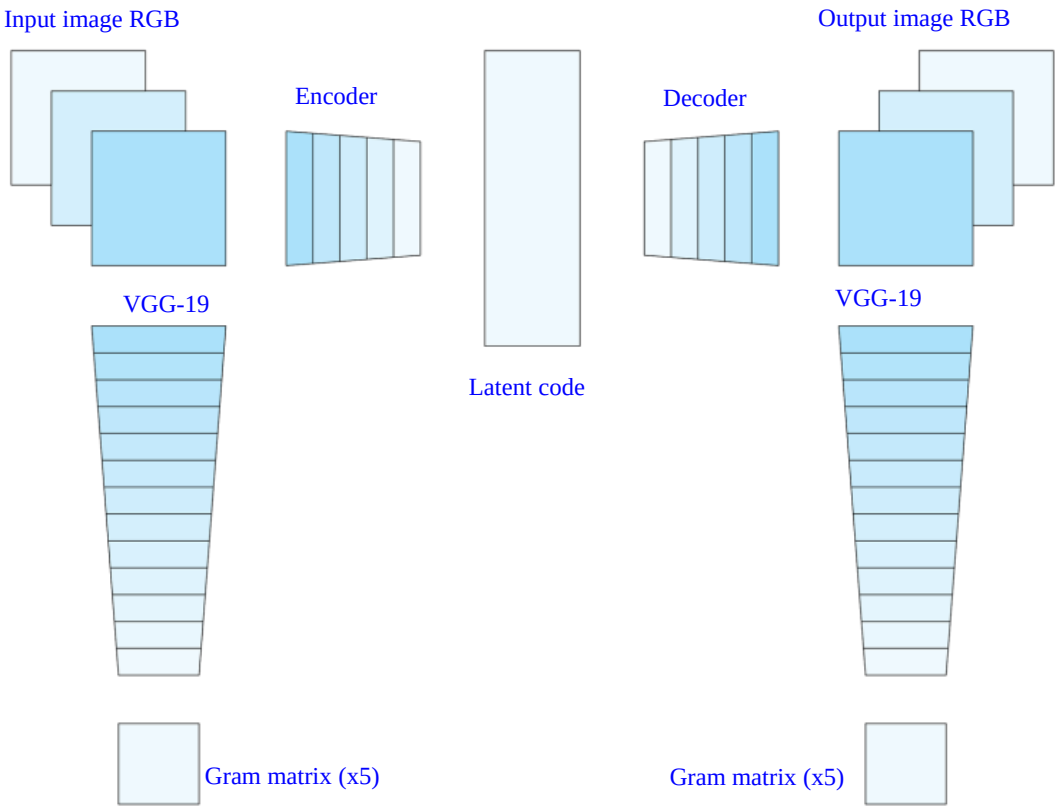


Figure 2.5: **Diagram of the model's architecture.** The backbone is a VAE (top), equipped with two VGG-19 instances (middle) to extract the Gram matrices (bottom).

VGG-19
Input image ($224 \times 224 \times 3$)
conv-64 conv-64
maxpool
conv-128 conv-128
maxpool
conv-256 conv-256 conv-256
maxpool
conv-512 conv-512 conv-512
maxpool
conv-512 conv-512 conv-512 conv-512
maxpool
fully connected 4096
fully connected 4096
fully connected 1000
softmax

Table 2.2: **Architecture of VGG-19.** Adapted from Simonyan and Zisserman (2015), Table 1.

2.6.2 Experimental configurations

	Dataset	MNIST	CIFAR-10
Parameters	Learning rate	10^{-4}	10^{-4}
	Channels	1	3
	Height	128	32
	Width	128	32
	Epochs	100	100
	Batch	60	60
Hyperparameters	β	1	1
	Latent dimension Z	10	10

Table 2.3: **Model parameters and hyperparameters for MNIST and CIFAR-10.** β is fixed at 1 so that the KL has the same weight as the reconstruction loss; Z is fixed to have the same value as the number of classes in the datasets.

	Dataset	custom dataset
Parameters	Learning rate	10^{-4}
	Channels	3
	Height	224
	Width	224
	Epochs	20000
	Batch	15
Hyperparameters	β	0.0005
	Latent dimension Z	20

Table 2.4: **Model parameters and hyperparameters for the custom dataset.** The values of the hyperparameters β and Z were chosen by performing a K-fold crossvalidation analysis, while the parameters were manually adjusted.

	Dataset	NSD
Parameters	Learning rate	10^{-3}
	Decay lr	10^{-4}
	Channels	3
	Height	224
	Width	224
	Epochs	300
	Days	20
	Decay day	18
	lr trans. day	4
	Batch	50
Hyperparameters	β	0.00052
	Latent dimension Z	36
	α	1-0

Table 2.5: **Model parameters and hyperparameters for NSD.** The values of the hyperparameters β and Z were chosen by performing a K-fold crossvalidation analysis; α value was adapted for standard interpolation pipeline; the parameters were manually adjusted.

	Model	AE	VAE
Parameters	Learning rate	10^{-4}	10^{-4}
	Channels	3	3
	Height	224	224
	Width	224	224
	Epochs	100	10.000
	Batch	8	8
Hyperparameters	β		0.0001
	Latent dimension Z		47

Table 2.6: **Model parameters and hyperparameters for DTD in the VAE training experiment.** The values β was manually set to be very small to purposely underweight the variational rule; Z dimensionality was kept equal to the number of semantic classes of DTD.

2.6.3 Supplementary figures

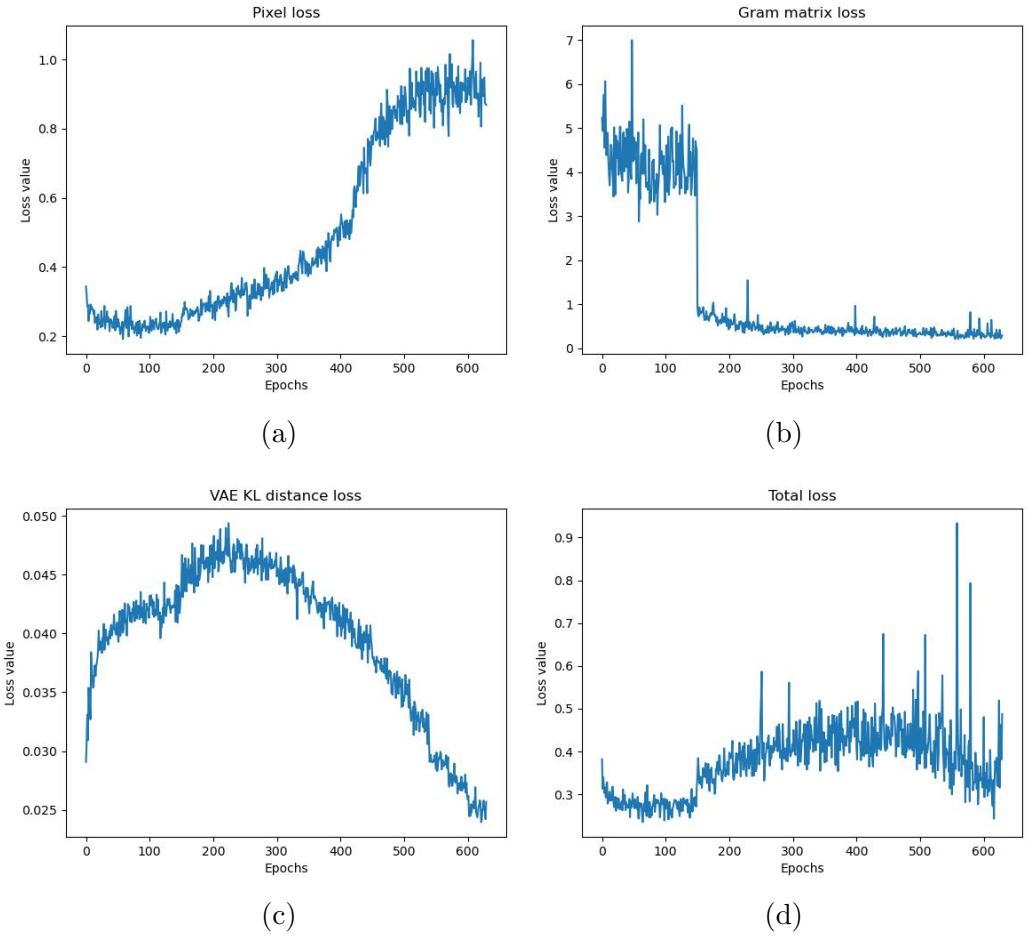
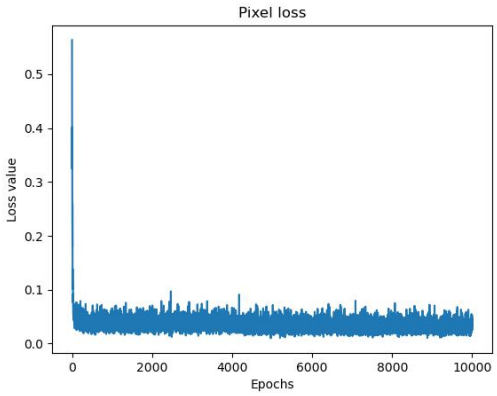
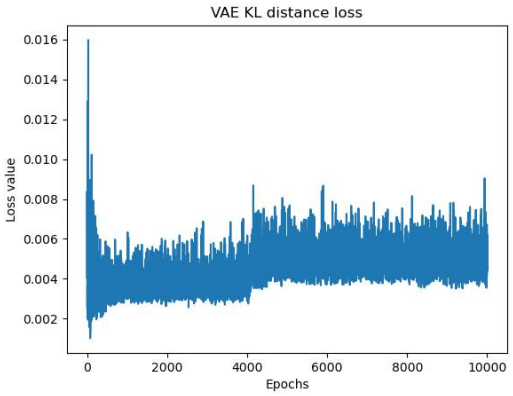


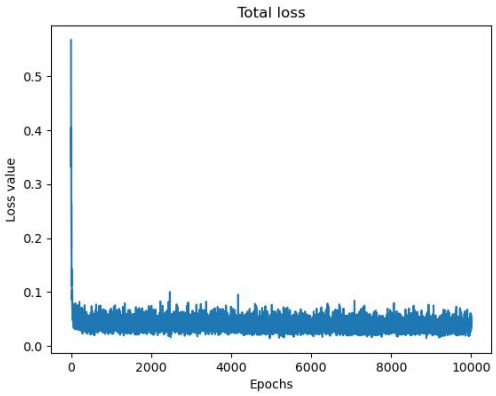
Figure 2.6: **Losses landscapes of our model trained with NSD** after 2 training days.



(a)



(b)



(c)

Figure 2.7: **Losses landscapes of VAE trained with DTD** at 10.000 epochs.

Bibliography

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1): 116–126, 2022. doi: 10.1038/s41593-021-00962-x.

Horace B. Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12(3):241–253, 2001. doi: 10.1088/0954-898X/12/3/301.

Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf. Learning texture manifolds with the periodic spatial GAN. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 469–477. PMLR, 2017.

Nalini Bhushan, A. Ravishankar Rao, and Gerald L. Lohse. The Texture Lexicon: Understanding the Categorization of Visual Texture Terms and Their Relationship to Texture Images. *Cognitive Science, 21*, pages 219–246, 1997.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 2095–2102, 2018.

Riccardo Caramellino, Eugenio Piasini, Andrea Buccellato, Anna Carboncino, Vijay Balasubramanian, and Davide Zoccolan. Rat sensitivity to multipoint statistics is predicted by efficient coding of natural scenes. *eLife*, 10:e72081, 2021. doi: 10.7554/eLife.72081.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Jeremy Freeman, Corey M. Ziemba, David J. Heeger, Eero P. Simoncelli, and J. Anthony Movshon. A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, 16(7):974–981, 2013. doi: 10.1038/nn.3402.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Ian Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2017.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014.

Shouchang Guo, Valentin Deschaintre, Douglas Noll, and Arthur Roullier. U-attention to textures: Hierarchical hourglass vision transformer for universal texture synthesis. In *Proceedings of the 19th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–10. Association for Computing Machinery, 2022. doi: 10.1145/3565516.3565525.

Eric Heitz, Kenneth Vanhoey, Thomas Chambon, and Laurent Belcour. A sliced wasserstein loss for neural texture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9412–9420, 2021. doi: 10.1109/CVPR46437.2021.00929.

Margaret M. Henderson, Michael J. Tarr, and Leila Wehbe. A texture statistics encoding model reveals hierarchical feature selectivity across human visual cortex. *Journal of Neuroscience*, 43(22):4144–4161, 2023. doi: 10.1523/JNEUROSCI.1822-22.2023.

Ann M. Hermundstad, John J. Briguglio, Mary M. Conte, Jonathan D. Victor, Vijay Balasubramanian, and Gašper Tkačik. Variance predicts salience in central sensory processing. *eLife*, 3:e03722, 2014. doi: 10.7554/eLife.03722.

Akshay V. Jagadeesh and Justin L. Gardner. Texture-like representation of objects in human visual cortex. *Nature Neuroscience*, 24:224–231, 2021. doi: 10.1038/s41593-020-00751-2.

Nikolay Jetchev, Urs Bergmann, and Roland Vollgraf. Texture synthesis with spatial generative adversarial networks. *arXiv preprint arXiv:1611.08207*, 2016.

Bela Julesz. Visual pattern discrimination. *IRE Transactions on Information Theory*, 8:84–92, 1962.

Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981. doi: 10.1038/290091a0.

Bela Julesz, E. N. Gilbert, L. A. Shepp, and H. L. Frisch. Inability of humans to discriminate between visual textures that agree in second-order statistics revisited. *Perception*, 2(4):391–405, 1973.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. The MNIST database of handwritten digits. https://yann.lecun.com/exdb/mnist/. Accessed: 2026-05-25.

Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000. doi: 10.1023/A:1026553619983.

Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in Neural Information Processing Systems*, volume 29, pages 2352–2360, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, June 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Omry Sendik and Daniel Cohen-Or. Deep correlations for texture synthesis. *ACM Transactions on Graphics*, 36(4):161:1–161:15, 2017. doi: 10.1145/3072959.3015461.

Eero P. Simoncelli. Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, 13(2):144–149, 2003. doi: 10.1016/S0959-4388(03)00047-3.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pages 1–14, 2015.

Xavier Snelgrove. High-resolution multi-scale neural texture synthesis. In *Proceedings of the SIGGRAPH Asia 2017 Technical Briefs*, SA '17. Association for Computing Machinery, 2017. doi: 10.1145/3145749.3149449.

P. Sterling and S. Laughlin. Principles of Neural Design . *The MIT Press*, 2017.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023. doi: 10.1109/CVPR52729.2023.01389.

Tiberiu Tesileanu, Mary M. Conte, John J. Briguglio, Ann M. Hermundstad, Jonathan D. Victor, and Vijay Balasubramanian. Efficient coding of natural scene statistics predicts discrimination thresholds for grayscale textures. *eLife*, 9:e54347, 2020. doi: 10.7554/eLife.54347.

Tiberiu Tesileanu, Eugenio Piasini, and Vijay Balasubramanian. Efficient processing of natural scenes in visual cortex. *Frontiers in Cellular Neuroscience*, 16:1006703, 2022. ISSN 1662-5102. doi: 10.3389/fncel.2022.1006703.

Gašper Tkačik, Jason S. Prentice, Jonathan D. Victor, and Vijay Balasubramanian. Local statistics in natural scenes predict the saliency of synthetic textures. *Proceedings of the National Academy of Sciences*, 107(42):18149–18154, 2010. doi: 10.1073/pnas.0914916107.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6924–6932, July 2017.

Jonathan Vacher, Aida Davila, Adam Kohn, and Ruben Coen-Cagli. Texture interpolation for probing visual perception. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22146–22157, 2020.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 6306–6315. Curran Associates, Inc., 2017.

J. H. van Hateren. A theory of maximizing sensory information. *Biological Cybernetics*, 68:23–29, 1992. doi: 10.1007/BF00203134.

J. D. Victor and M. M. Conte. Local image statistics: maximum entropy constructions and perceptual salience. *Optical Society of America*, 29:1313–1345, 2012.

Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, volume 9911 of *Lecture Notes in Computer Science*, pages 835–851. Springer, 2016. doi: 10.1007/978-3-319-46478-7_51.

Thomas S. A. Wallis, Christima M. Funke, Alexander S. Ecker, Leon A. Gatys, Felix A. Wichmann, and Mathias Bethge. A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of Vision*, 17: 1–29, 2017.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In *International Conference on Learning Representations*, 2022.

Mirko Zanon, Bastien S. Lemaire, Eugenio Piasini, Riccardo Caramellino, Caroline Nallet, Vijay Balasubramanian, Judit Gervain, Davide Zoccolan, and Giorgio Vallortigara. Predisposed and learned preferences for multipoint visual statistics in visually naïve newly hatched chicks. *bioRxiv*, 2025. bioRxiv preprint.

CHAPTER 3

Perceptual misalignment of texture representations in convolutional neural networks

This chapter is a reproduction of the paper Perceptual misalignment of texture representations in convolutional neural networks, currently accepted in *Journal of Vision*.

Statement of contribution: I performed all experiments and analyses, made all the figures, wrote the first draft of the paper, and participated in further editing of the draft and in responding to peer review. Fabio Anselmi, Alessio Ansuini and Eugenio Piasini provided scientific guidance, contributed to the conceptual framing and discussion of the findings, and participated in editing the paper and responding to peer review.

3.1 Abstract

Mathematical modeling of visual textures traces back to Julesz's intuition that texture perception in humans is based on local correlations between image features (Julesz, 1981). An influential approach for texture analysis and generation generalizes this notion to linear correlations between the nonlinear features computed by convolutional neural networks (CNNs), compiled into Gram matrices (Gatys et al., 2015). Given that CNNs are often used as models for the visual system, it is natural to ask whether such "texture representations" spontaneously align with the textures' perceptual content, and in particular whether those CNNs that are regarded as better models for the visual system also possess more human-like texture representations. Here we quantify the perceptual content captured by feature correlations computed for a diverse pool of CNNs, and we compare it to the models' perceptual alignment with the mammalian visual system as measured by Brain-Score (Schrimpf et al., 2018). Surprisingly, we find that there is no connection between conventional measures of CNN quality as a model of the visual system and its alignment with human texture perception. We conclude that texture perception involves mechanisms that are distinct from those that are commonly modeled using approaches based on CNNs trained on object recognition, possibly depending on the integration of contextual information.

3.2 Introduction

Visual textures hold a special status in vision neuroscience as stimuli that afford a non-trivial degree of controllability and complexity, unlike natural images or more traditional parametric stimuli (Victor et al., 2017). Because of this, multiple mathematical frameworks for modeling and generating textures have been established, with considerable success. One particularly influential strand of theoretical, computational and empirical work stems from the seminal view of (Caelli and Julesz, 1978; Julesz, 1962, 1981), who proposed that texture perception is controlled by local correlations of visual features within an image. This view was extended by later work that better formalized the notion of

textures as statistical objects (Zhu et al., 2000) (but see Liu et al. (2019) for a detailed historical overview). This study demonstrated the effectiveness of using features based on the neuroscience of vision (Portilla and Simoncelli, 2000) and developed a connection with natural stimulus statistics through the efficient coding paradigm (Caramellino et al., 2021; Hermundstad et al., 2014; Tesileanu et al., 2020, 2022; Tkačik et al., 2010; Victor and Conte, 2012; Victor et al., 2017; Zanon et al., 2025).

In recent years, following the rise in popularity of deep convolutional neural networks (CNNs) as tools in computer vision, Gatys et al. proposed a machine learning-based operationalization of Julesz's idea (Gatys et al., 2015). The Gatys model effectively defined a texture ensemble based on the Gram matrices formed by spatial correlations between intermediate-layer activations of a CNN trained for object recognition. Texture synthesis is performed by optimizing the image pixels via gradient descent to match the statistical structure of an exemplar texture. This is achieved by minimizing a *Gram loss*, defined as the normalized L2 distance between the Gram matrices of the exemplar image and those of the synthesized image. This approach was rapidly adopted in the budding fields of neural texture synthesis and style transfer (see for instance Gatys et al. (2016); Johnson et al. (2016); Li et al. (2017); Ulyanov et al. (2017); Yu et al. (2019)), and was later improved under multiple aspects: high resolution images processing (**?**), better methods for capturing long-range correlations within images (Berger and Memisevic, 2017; Liu et al., 2016; **?**), usage of more sophisticated losses capable of capturing stationary properties of the CNN's activations beyond the pairwise correlations in the Gram loss (Heitz et al., 2021; Vacher et al., 2020).

In parallel, and more broadly, CNNs trained for object recognition have become ubiquitous tools in vision neuroscience. Over the years, work from many labs has established these models as functional accounts of the ventral visual stream in monkeys (Bashivan et al., 2019; Cadena et al., 2019; Cadieu et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Schrimpf et al., 2020; Yamins et al., 2014) and humans (Güçlü and van Gerven, 2015; Storrs et al., 2021), as well as the corresponding visual cortical areas in rodents (Cadena et al., 2019; Muratore et al., 2022, 2025; Nayebi et al., 2023). However, the strong predictive power of this approach for neural activity is only partly mirrored by the capacity of these artificial neural networks to accurately predict behavioral responses to individual stimuli (Feather et al., 2023; Geirhos et al., 2021; Rajalingham et al., 2018; Vinken et al., 2020; Wichmann and Geirhos, 2023).

Beyond object recognition, the ventral stream is also thought to be important for texture perception. Indeed, the idea that texture perception can be explained by reference to the neural mechanisms in early vision goes at least as far back as Turner (1986) and Malik and Perona (1990). More recent results in computational modeling and neurophysiology evidence point generally to the involvement of early-to-intermediate areas in the ventral stream (Freeman et al., 2013; Henderson et al., 2023; Matteucci et al., 2024; Okazawa et al., 2017; Portilla and Simoncelli, 2000; Yu et al., 2015; Ziemba et al., 2016, 2024). However, overall, our understanding of the neural computations underpinning texture perception is still limited compared to those enabling object recognition. This makes CNN-based methods for texture analysis and synthesis particularly interesting from a neuroscience perspective, despite the recent emergence of approaches based on diffusion models for practical applications, e.g. (Wang et al., 2024). More specifically, it raises the question of whether the existing quality measures devised for CNN-based models of the ventral stream align with the perceptual quality of textures generated based on the internal features of those models. If such alignment exists, it would not only suggest that object recognition and texture perception share a significant amount of functional mech-

anisms, but also that these mechanisms are among those that are well captured by CNNs trained on object recognition. Conversely, the absence of such a relationship – where increased fidelity to neural or behavioral object recognition data does not lead to improved modeling of human texture perception – would indicate that current CNN architectures or training paradigms lack critical components.

In the present work, we considered a set of popular CNNs and we asked whether the models that are more "brain-like" in object recognition also tend to represent visual textures in a way that is closer to human perception. As a metric of model performance on object recognition, we used the standardized benchmarks provided by Brain-Score (Schrimpf et al., 2018). To measure alignment with human texture perception, we evaluated the models on the images of the Describable Texture Dataset (Cimpoi et al., 2014), which are divided in classes according to human annotations. We then assessed whether the "texture representations" provided by the Gram matrices computed from the models' feature maps (as in Gatys et al. (2015)) are organized in a way that respect the classes subdivision in the dataset.

Our results show that there is no correlation between the goodness of the Gram texture representation – with respect to the clusters formed by human-annotated labels – and Brain-Score. This suggests that the features governing the utility of artificial neural networks as models of visual cortex in an object recognition context are not fully aligned with those that can make them good models for texture processing, therefore highlighting an underexplored path for the improvement of machine learning-based models for texture analysis and synthesis.

3.3 Methods

3.3.1 The "Gram representation" for textures

Our study builds on the texture analysis and synthesis method introduced by Gatys et al. (Gatys et al., 2015). This approach models textures using feature representations extracted from a CNN, which are shown to support the generation of perceptually compelling texture samples. In the original implementation, given a natural texture image x, the algorithm starts by feeding x as input to the network and extracting the corresponding activation maps from a predetermined set of layers. For each selected layer l, the feature activations F^l are used to compute a Gram matrix G^l, defined as the inner product between vector representations of the feature maps:

$$G^l_{ij} = \sum_m F^l_{im} F^l_{jm}, \tag{3.1}$$

where the index m runs over spatial samples of the feature maps, and i and j label feature channels at layer l. This Gram matrix summarizes the correlations between feature channels, discarding spatial information and yielding a statistical description of textures. These "Gram representations" are thought to capture the elements of the input image that are important for texture perception. In this work, we study the Gram representations generated by a number of different CNNs.

3.3.2 Texture synthesis

In the present work we provide examples of texture synthesis to help the reader gauge the quality of the images that can be produced based on the representations described in

the previous paragraph, despite the fact that our main analyses and conclusions do not depend on the output of any texture synthesis algorithm. For completeness, we report here the details of our synthesis approach.

To generate synthetic textures we followed the method established by Gatys et al. (2015), with minor modifications. Briefly, texture synthesis is performed by iteratively updating the pixels of an initially random image $\hat{x}$, where pixel values are drawn from a Gaussian distribution $\hat{x} \sim \mathcal{N}(0, 1)$. This process leads to match the Gram matrices of $\hat{x}$ to those of the target image x by minimizing the *Gram loss*, defined as a weighted sum of squared L2 distances between corresponding Gram matrices across layers:

$$\mathcal{L}(\hat{x}, x) = \sum_l \frac{1}{4M_l N_l^2} \|\hat{G}^l - G^l\|_2^2. \tag{3.2}$$

where N_l is the number of features in layer l and M_l is the number of units within each feature (note that the term $1/4M_l N_l^2$ that we use here to weight the contribution from each individual layer is not identical to the one in Gatys et al. (2015); see the Appendix for details on our choice of loss weighting for image synthesis). To minimize the Gram loss we used L-BFGS as in Gatys et al. (2015), using the implementation of the optimization algorithm provided by PyTorch (Ansel et al., 2024).

Minimizing this objective in pixel space generates novel images whose texture statistics, as captured by feature correlations in a convolutional neural network, match those of the original image. Importantly, while spatial structure is not preserved, the resulting images retain perceptually salient texture properties, highlighting the role of feature-level statistics in texture representation.

3.3.3 CNN Layer selection

We implemented the texture representation and the synthesis algorithm described above using PyTorch (Ansel et al., 2024), for a range of CNN architectures trained for object classification with ImageNet-1K (see Table 3.1). The pre-trained networks were sourced from TorchVision (TorchVision maintainers and contributors, 2016). The convolutional layers of the CNN originally used by Gatys et al. (2015) (VGG-19) have five different spatial scales, corresponding to the five consecutive "convolve-and-pool" blocks of that architecture. Gatys et al. (2015) showed that selecting just one layer for each of these spatial scales was enough to produce perceptually satisfying synthetic textures. Therefore to ensure that the results of our analysis are comparable across different networks, we extracted feature maps from five layers from each architecture, selected in such a way that they are located at comparable relative depth across architectures; see Appendix, Supplementary Tables 3.4, 3.5 for details.

3.3.4 Stimulus dataset

The Describable Textures Dataset (DTD) (Cimpoi et al., 2014) is our input dataset of choice. It consists of 5640 images of naturalistic visual textures collected from the Internet, spanning a wide range of materials and surface appearances. The dataset is organized into 47 texture classes, each one containing 120 images. The images size ranges between 300x300px and 640x640px. We resized (scaled) every image to 224x224px, in order to match the CNNs' input spatial dimensions and to maintain a stable representation of the textures.

Family	Model
	AlexNet (**?**)
DenseNet (Huang et al., 2017)	DenseNet-121 DenseNet-169 DenseNet-201
InceptionNet (Howard et al., 2017)	Inception-v3 (Szegedy et al., 2016)
ResNet (He et al., 2016)	ResNet18 ResNet34 ResNet50 ResNet101 ResNet152
VGG (Simonyan and Zisserman, 2015)	VGG-16 VGG-19

Table 3.1: **Convolutional Neural Network architectures grouped by family.**

The class labels arise from a human annotation procedure based on the *Texture Lexicon*, a psycholinguistic study introduced in Bhushan et al. (1997) with the goal of constructing a linguistically grounded vocabulary for describing visual textures. The Texture Lexicon identifies a set of texture-related adjectives commonly used by human observers (e.g., *blotchy*, *striped*, *bumpy*, *woven*) and organizes them according to perceptual relevance.

To construct DTD, the authors selected 47 representative texture terms from the Texture Lexicon and used them as labels in a large-scale annotation task involving human participants. Annotators were presented with images and asked to assign one or more texture attributes that best described the visual appearance of the image. Final class labels were determined by aggregating human responses and retaining images for which there was sufficient agreement across annotators. As a result, DTD encodes texture categories that reflect shared perceptual judgments rather than low-level image statistics or semantic object labels.

This annotation procedure makes DTD particularly suitable for studies of texture perception, as it provides a direct link between image-level texture statistics and perceptually meaningful texture descriptors grounded in human language and cognition. DTD treats textures as entities that can be categorized in a manner analogous to object images, effectively providing human-defined clusters of visual textures with a balanced distribution of images across classes. The richness and diversity of DTD thus constitute a perceptual and linguistic benchmark against which we can evaluate whether texture representations derived from CNNs in combination with the Gatys method are sufficient to preserve class-specific texture distinctions as defined by human perception.

3.3.5 Clustering

We performed Representation Similarity Analysis (RSA) (Kriegeskorte et al., 2008) on the texture representations extracted with the Gram matrices on our pool of 12 CNNs. RSA was introduced in neuroscience as a method to compare object representations obtained with different means (e.g., functional Magnetic Resonance Imaging vs. Magnetoencephalography), or to compare biological and computational representations (e.g., brain activity vs. CNNs). It relies on Representational Dissimilarity Matrices (RDMs), where each entry is the pairwise dissimilarity between responses to two stimuli. We employed

RSA to verify whether the RDMs produced with Gram matrices extracted from DTD show a clustering pattern consistent with the original 47 classes. For each one of the 12 CNNs, we obtained 5 RDMs – one for each of the convolutional layers analyzed – of dimensionality 5640x5640, where each entry is the distance among Gram matrices for one pair of images in DTD.

We tested whether Gram matrix–based texture representations capture the structure of the 47 texture classes in DTD by performing hierarchical clustering with Ward linkage function based on Euclidean distance, as implemented in scikit-learn (Pedregosa et al., 2011). Hierarchical clustering is an unsupervised technique that groups data points based on their similarity, producing a tree-like representation in which clusters are progressively merged according to their pairwise distances. This hierarchical organization provides insight into the intrinsic structure of the data without relying on class labels.

To evaluate the extent to which the learned texture representations recover the human-defined categories in DTD, we treated the dataset labels as a reference standard. Specifically, we fixed the number of clusters to $k = 47$, matching the number of texture classes in the original dataset, and allowed the clustering algorithm to organize the representations accordingly. This procedure enables a direct comparison between the automatically discovered clusters and the perceptual texture classes defined by human annotations.

3.3.6 Measuring clustering quality

We quantified texture information carried in the representations within Gram matrices with Mutual Information (MI) (MacKay, 2003). In our analysis MI was computed between the real 47 classes in DTD (*real classes*) and the 47 clusters found by hierarchical clustering (*found clusters*). If X and Y are random variables representing respectively the true label of an image (*real class*) and the label assigned to it by the clustering procedure (*found cluster*), MI is the reduction in uncertainty, quantified as entropy (H), in one of the two variables associated with learning the value of other:

$$\mathrm{MI}(X:Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = \sum_{x,y} p(x,y)\, \log \frac{p(x,y)}{p(x)\, p(y)}$$

By construction, $\mathrm{MI}(X:Y)$ is upper-bounded by the minimum between $H(X)$ and $H(Y)$. This can be seen by the expression above taking into account that $H(X|Y)$ and $H(Y|X)$ are always nonnegative. Therefore, considering that Y follows an unknown distribution where *found clusters* result from hierarchical clustering, and X is distributed uniformly among the 47 *real classes* of DTD, in our case the MI is upper bounded by the entropy of X:

$$\mathrm{MI}_{\max} = H(X) = -\sum_{i=1}^{47} p_i \log_2 p_i = \log_2 47 \approx 5.5\text{bits} \tag{3.3}$$

All empirical estimates for mutual information were corrected for potential undersampling bias using the NSB method (Nemenman et al., 2002), as implemented by the ndd Python package (Marsili, 2021).

We quantified MI between *real classes* and *found clusters* for the *found classes* distribution drawn from each of the 5 layers of the 12 CNNs. Therefore, for each model we obtained 5 MI values.

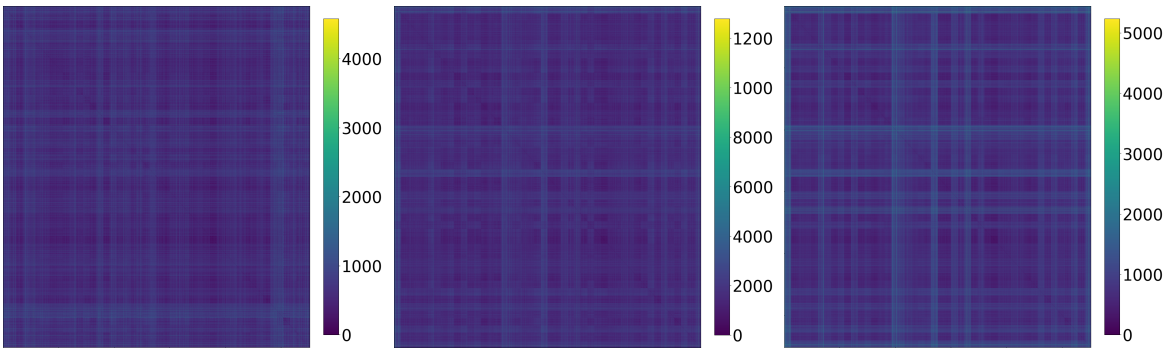
3.3.7 Brain-Score correlation

To assess the relationship between the goodness of a CNN as a model for biological object recognition and texture perception, we correlated the best MI value (i.e. the highest out of the 5 MI values computed for each model) of each CNN to Brain-Score. Brain-Score is a composite of multiple behavioral and neural benchmarks that determine how similar artificial neural networks are to the object recognition mechanisms in the human and macaque brain (Schrimpf et al., 2018). In this study, we focus on the following Brain-Score components: *neural vision*, *behavior vision*, and *average vision*, as well as the region-specific neural predictability scores for *V1, V2, V4*, and *IT*. The *neural vision* score summarizes a model's ability to predict neural responses across areas of the primate ventral visual stream, while *behavior vision* quantifies the alignment between model predictions and human behavioral performance on object recognition tasks. The *average vision* score provides an overall summary across neural and behavioral benchmarks. The region-specific scores (*V1, V2, V4, IT*) evaluate how well model representations predict neural activity in individual cortical areas involved in the hierarchical processing of visual information along the ventral stream. For each of the 12 models, we correlated the highest MI with the aforementioned Brain-Score values which we report in Appendix Supplementary Table 3.6. We run a linear correlation with Pearson's p and set $\alpha = 0.05$.

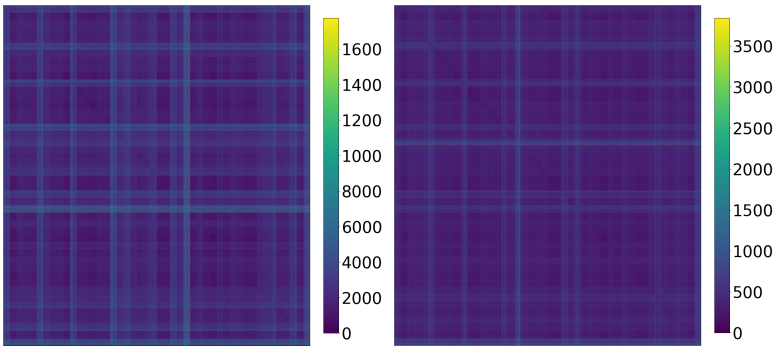
3.4 Results

3.4.1 Representational Similarity Analysis

To understand how texture information is represented across the layers of our CNNs, we started by performing Representational Similarity Analysis. Figure 3.1 shows Representational Dissimilarity matrices (RDMs) calculated on pairs of Gram matrix representations of DTD images extracted from VGG-19. Performing the same analysis on other networks led to qualitatively similar results (see de Paolis (2025) for high-resolution RDM plots for all the networks considered in this work). Visual inspection reveals that the structure of the RDMs reflects the existence of at least some of the classes in DTD, and that for this network the clearest structure emerges in the last three layers. To understand these trends more quantitatively, we next moved to performing clustering analysis of the Gram representations.



Extracted layer n. 1 Extracted layer n. 2 Extracted layer n. 3



Extracted layer n. 4 Extracted layer n. 5

Figure 3.1: **Representational Dissimilarity Matrices** obtained from the five layers analyzed for VGG-19. The size of each matrix is 5640×5640, and each entry corresponds to the Euclidean distance – relative to each matrix, as in the colorbars – between the Gram matrix representation of one pair of images in DTD. The entries are sorted such that images belonging to each of the 47 ground-truth categories are placed next to each other.

3.4.2 Clustering analysis

Figure 3.2 represents the values of MI by model and by layer found with hierarchical clustering. The plot shows a trend: MI grows along with the depth of the extracted features, consistently with the trend found by visual inspection in the RSA plots. The highest MI is yielded by the last layer for most models, but it does not go beyond ≈ 2.7 bits, which is about half of the maximum achievable MI, according to Equation 3.3. These numbers show that increasing the depth of the extracted representations leads to an increasing clustering quality.

The results of the correlation among the highest MI value per model and selected Brain-Score benchmarks are reported in Table 3.2. We did not find any significant correlation between any Brain-Score value and the best MI found in any model, as shown in Figure 3.3.

Brain-Score metric	**Pearson's r**	**p-value**
average vision	-0.130	0.68
neural vision	-0.165	0.60
behavior vision	-0.029	0.92
V1	-0.229	0.47
V2	-0.262	0.41
V4	-0.063	0.84
IT	-0.104	0.74

Table 3.2: **Correlation between texture representation quality and BrainScore.** Texture representation quality is expressed as MI between the clusters in the Gram representation of textures and the ground-truth labels in DTD, and the quality of the networks as models of the ventral visual stream is measured by Brain-Score. For more details on Brain-Score see Section 3.3.7.

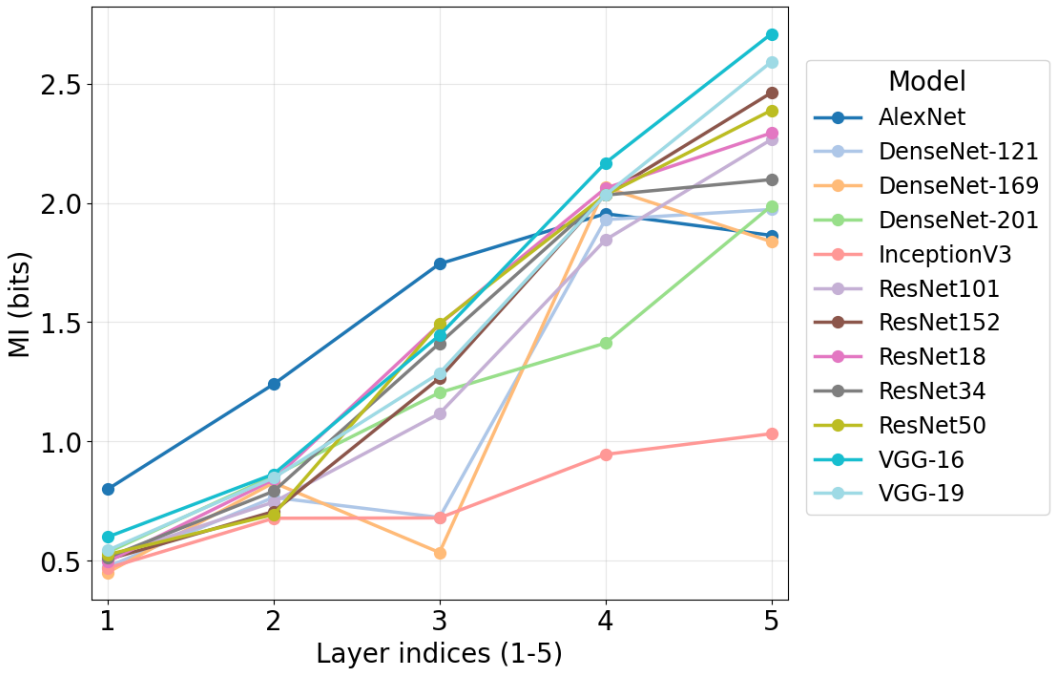


Figure 3.2: **MI values across CNN layers** identified by index (1-5). Each line corresponds to one of the 12 models as color-coded in the legend.

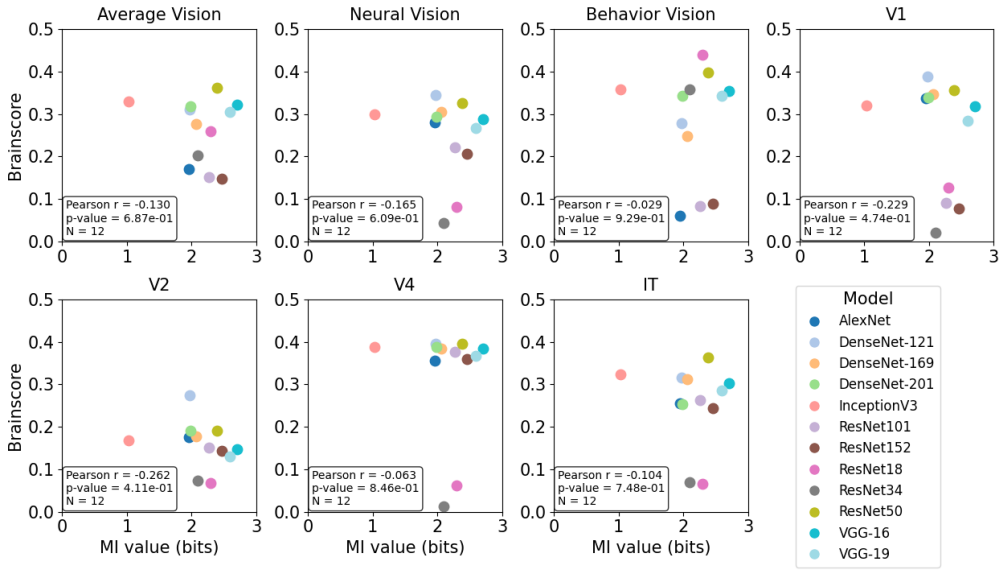


Figure 3.3: **Plots of the correlations with Pearson's *p*.** Each subplot represents the correlation of Brain-Score benchmarks *average vision, neural vision, behavior vision, V1, V2, V4, IT* respectively vs. the layer with the highest MI for each of the 12 CNNs. Each color identifies a model as reported in the legend.

3.4.3 Texture synthesis

In Figure 3.3 we report samples of synthesized textures for a subset of selected models. The purpose of this plot is to illustrate the output of the Gatys algorithm, as implemented by us with our pool of CNNs; we underline that texture synthesis is not part of the main analysis pipeline, and in our work is only used to generate example visualizations. The results of the quantitative analyses described in the previous sections do not depend directly from this visual outcome. Each sample belongs to 4 separate DTD classes: *blotchy, striped, matted, scaly.* We observe how some models achieve a better perceptual output in terms of synthesized textures, despite running the same algorithm.

3.5 Discussion

In this study, we have investigated the quality of the Gram texture representation proposed by Gatys et al. (2015) in terms of its agreement with the categorical grouping in the Describable Texture Dataset (Cimpoi et al., 2014), for a diverse pool of deep convolutional networks. We have compared this measure of perceptual quality with Brain-Score, an independent benchmark for "brain-like" operation of the CNNs in object recognition settings (Schrimpf et al., 2018, 2020).

Our analyses lead to three main conclusions. First, the perceptual quality of the Gram representations tends, for most networks, to grow monotonically with the depth of the layers analyzed (Figure 3.2). This is generally compatible with recent results suggesting that texture processing is distributed along intermediate areas of the ventral visual stream inclusing V2 and V4 (Okazawa et al., 2017; Ziemba et al., 2024) and possibly IT (Zhivago and Arun, 2014).

Second, the Gram representation is unable to recover the full class structure of the

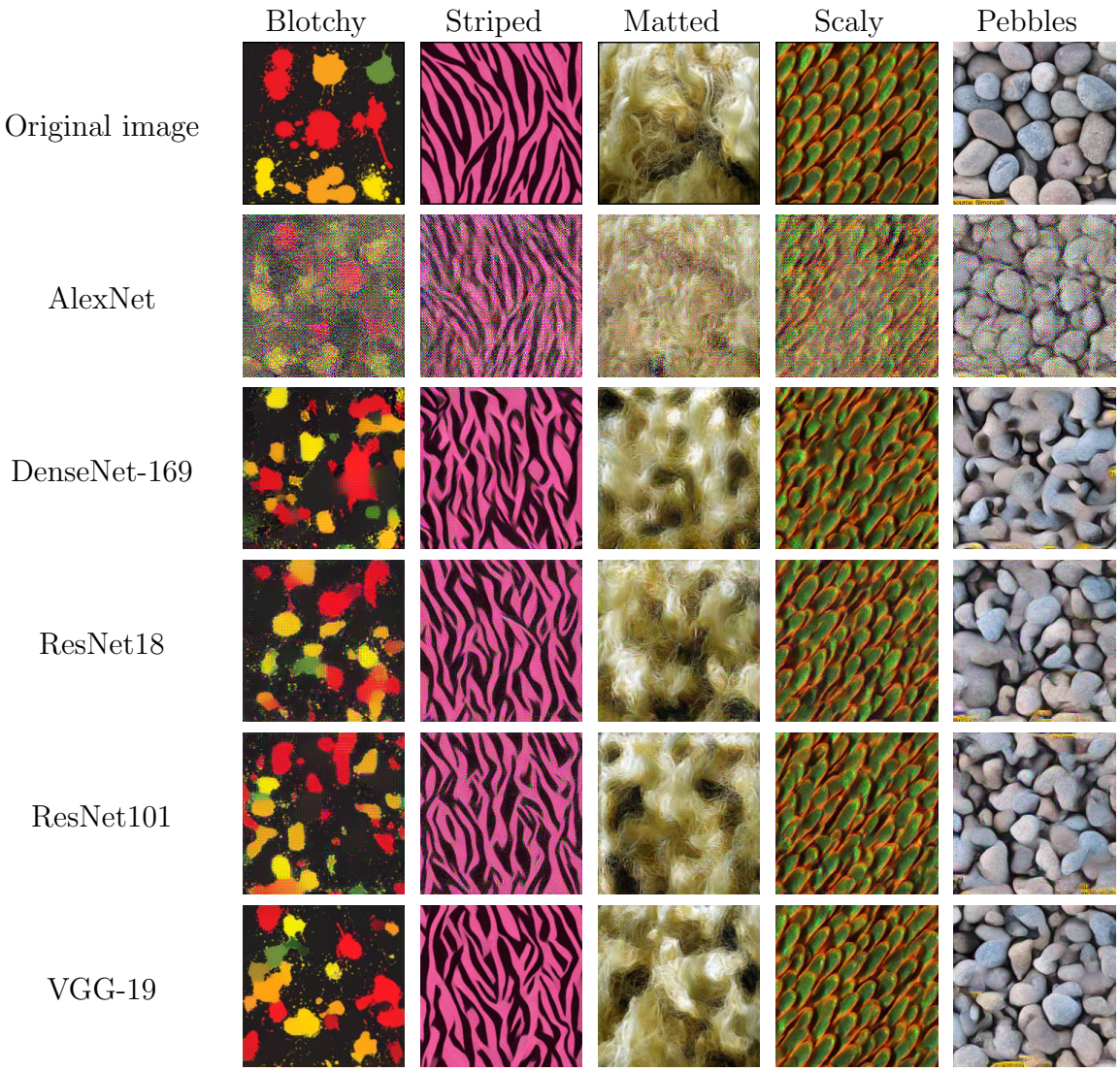


Table 3.3: **Textures generated with the Gatys algorithm** applied to a subsample of CNNs.

data, as determined by the human-assigned labels. Quantitatively, this is seen by the value of the clustering MI reaching only about half of the theoretical maximum (Figure 3.2). This highlights that, despite its popularity, the Gram representation is far from perfect. In further work, we aim at exploring whether the perceptual improvements associated with more sophisticated versions of the representation translate to significant improvements in MI on the DTD.

Third, the quality of the CNNs as implicit models for texture perception is not correlated with their quality as models of the ventral stream in an object recognition setting (Figure 3.3). This is despite the fact that there is substantial variability in both metrics, as is well known for Brain-Score, and as can be also appreciated visually for the textures in Figure 3.3. It is not simply due for instance to all models having roughly the same texture representation capacity. This result is surprising, because one could have expected that the functional features captured by Brain-Score are general enough to at least partly include those that are necessary to support human-like texture perception. Instead, this appears not to be the case. We conclude that in order to design methods of texture analysis and synthesis that are better aligned with human perception we need to look beyond standard CNNs trained for object recognition, for instance by incorporating unsupervised or self-supervised learning (Matteucci et al., 2024), multimodality (Radford et al., 2021) or different architectures, such as those with attention mechanisms (Guo et al., 2022; Vaswani et al., 2017).

Finally, we note that our conclusions are broadly compatible with another series of works relating human vision and texture encoding in CNNs. In an influential paper, Geirhos et al. (2019) showed that CNNs trained for object classification on ImageNet tend to classify images based on texture information more than shape information, in a way that differs markedly from human behavior. However, subsequent work within that research strand points to the idea that, in fact, shape and texture information do coexist within the representation layers of CNNs, and that this so-called "texture bias" emerges primarily at the level of the decision head, which under standard training conditions learns to prioritize texture information over shape information (Hermann et al., 2020; Islam et al., 2021). Our notion of "texture information" (the MI between the Gram clusters and the ground-truth DTD labels) increases with layer depth across most models, generalizing the observations in Hermann et al. (2020), which shows (Supplementary Figure C.4) that for AlexNet the texture information increases rapidly in early convolutional layers and then flattens out towards the end of the hierarchy. This is very similar to the trend we report for AlexNet in Figure 3.2. Another interesting observation in Hermann et al. (2020) is that neurally-motivated architectures like CorNET (Kubilius et al. (2018); designed by the same team as Brain-Score) do not fare better than standard CNNs from the point of view of their texture bias in classification. This echoes our own results showing that architectures that are better models of ventral stream processing in the canonical sense (related to object identity processing, measured by Brain-Score) are not necessarily more human-like in how they process textures.

3.6 Conclusions

Our work highlights a number of limitations of popular machine learning-based approaches for visual texture analysis and synthesis. At the same time, it provides a simple example of a visual function that is likely performed in the ventral stream but for which the existing quality metrics of neural and behavioral fidelity for models seem to have very little predictive power. In future work, we plan to extend our analyses to different architectures

and texture representations, and to develop a more direct and systematic way of testing the perceptual quality of generated textures with large-scale human psychophysics. Going forward, we hope that our approach can provide an additional guide for the development of truly general-purpose models of the mammalian visual system.

3.7 Appendix

3.7.1 Loss normalization

In this section we derive the weighting rule we used to combine the Gram losses from multiple layers.

Consider the Gram matrix G for a certain layer l. Its (i, j)-th entry is

$$G_{ij} = \sum_{n=1}^{M} F_i^n F_j^n$$

where the index n runs over all elements of the i-th feature map F_i (so for instance in a typical convolutional network M would be the product of the height and width of the feature map). Assume that the individual activations of the network units are roughly normally distributed:

$$F_i^n \sim \mathcal{N}(0, 1)$$

This is a reasonable assumption in presence of batch normalization. Our goal is now to understand how the "Gram loss" for this layer is distributed. The Gram loss is

$$\mathcal{L}_l = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(G_{ij}^{\text{rec}} - G_{ij}^{\text{orig}} \right)^2$$

For a fixed value of i, j and n let's define two new random variables A and B:

$$A = \left(\frac{F_i^n + F_j^n}{\sqrt{2}} \right)^2 \quad , \quad B = \left(\frac{F_i^n - F_j^n}{\sqrt{2}} \right)^2$$

from which we get

$$\frac{A - B}{2} = \frac{1}{2} \left[\frac{F_i^2 + F_j^2 + 2F_iF_j}{2} - \frac{F_i^2 + F_j^2 - 2F_iF_j}{2} \right] = \frac{1}{4} \left[4F_iFj \right] = F_iF_j$$

so we can think of G_{ij} as the sum of M iid terms, each of which is distributed like $(A-B)/2$. By observing that $(F_i^n + F_j^n)/\sqrt{2} \sim \mathcal{N}(0, 1)$, we see that A and B are each simply the square of a standard Normal. Therefore they both follow a Chi-squared distribution with one degree of freedom

$$A \sim \chi_1^2 \quad , \quad B \sim \chi_1^2$$

which means that

$$\mathbb{E}[A] = \mathbb{E}[B] = 1 \quad , \quad \text{Var}[A] = \text{Var}[B] = 2$$

Therefore, $A - B$ is the difference of two independent random variables with mean zero and variance 2. This implies that, for any n,

$$\mathbb{E}[2F_i^n F_j^n] = \mathbb{E}[A - B] = 0$$
$$\text{Var}[2F_i^n F_j^n] = \text{Var}[A] + \text{Var}[B] = 4$$

and therefore

$$\mathbb{E}[F_i^n F_j^n] = 0$$
$$\mathrm{Var}[F_i^n F_j^n] = 2$$

So G_{ij} is the sum of a large number M of iid random variables, each with mean 0 and variance 2. By the central limit theorem, then,

$$G_{ij} \sim \mathcal{N}\left(0, \sqrt{2M}\right)$$

If this is true for both the original and reconstructed Gram, then

$$G_{ij}^{\text{rec}} - G_{ij}^{\text{orig}} \sim \mathcal{N}(0, 2\sqrt{M})$$

and therefore

$$\frac{G_{ij}^{\text{rec}} - G_{ij}^{\text{orig}}}{2\sqrt{M}} \sim \mathcal{N}(0, 1)$$

Now, consider the quantity

$$\frac{\mathcal{L}_l}{4M} = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{G_{ij}^{\text{rec}} - G_{ij}^{\text{orig}}}{2\sqrt{M}}\right)^2$$

It is clear that $\mathcal{L}_l/4M$ is the sum of N^2 squared standard Normal random variables. Therefore,

$$\frac{\mathcal{L}_l}{4M} \sim \chi^2_{N^2}$$

and as a consequence

$$\mathbb{E}\left[\frac{\mathcal{L}_l}{4M}\right] = N^2$$

Finally, this implies that

$$\mathbb{E}\left[\mathcal{L}^l\right] = 4MN^2$$

In other words, if the activations F_i^n are standardized, the expected value of the Gram loss for a layer with N features of size M is $4MN^2$, just as a consequence of the size and shape of the layer. Therefore, to make sure that Gram losses coming from different layers have the same weight in our total loss we should divide each layer-specific Gram loss by $4MN^2$, where M and N are the values specific to that layer.

3.7.2 Supplementary Tables

Table 3.4: **Extracted layers from each CNN** and relative Gram matrix dimensionality. The names refer to the terminology of the models' versions in Torchvision. For models whose layers are not associated with a name, we report the absolute index.

Model	**Extracted layers**	**Gram dimensionality**
AlexNet	`layer idx=[0, 3, 6, 8, 10]`	L0=$64(64+1)/2 = 2080$ L3=$192(192+1)/2 = 18520$ L6=$384(384+1)/2 = 73920$ L8=$256(256+1)/2 = 32896$ L10=$256(256+1)/2 = 32896$

DenseNet-121	`features.norm0`
	`features.denseblock1.denselayer1.norm2`
	`features.denseblock2.denselayer1.norm2`
	`features.denseblock3.denselayer1.norm2`
	`features.denseblock4.denselayer1.norm2`
		bn0=$64(64+1)/2 = 2080$
		bn1=$128(128+1)/2 = 8256$
		bn2=$128(128+1)/2 = 8256$
		bn3=$128(128+1)/2 = 8256$
		bn4=$128(128+1)/2 = 8256$

DenseNet-169	`features.norm0`
	`features.denseblock1.denselayer1.norm2`
	`features.denseblock2.denselayer1.norm2`
	`features.denseblock3.denselayer1.norm2`
	`features.denseblock4.denselayer1.norm2`
		bn0=$64(64+1)/2 = 2080$
		bn1=$128(128+1)/2 = 8256$
		bn2=$128(128+1)/2 = 8256$
		bn3=$128(128+1)/2 = 8256$
		bn4=$128(128+1)/2 = 8256$

DenseNet-201	`features.norm0`
	`features.denseblock1.denselayer1.norm2`
	`features.denseblock2.denselayer1.norm2`
	`features.denseblock3.denselayer1.norm2`
	`features.denseblock4.denselayer1.norm2`
		bn0=$64(64+1)/2 = 2080$
		bn1=$128(128+1)/2 = 8256$
		bn2=$128(128+1)/2 = 8256$
		bn3=$128(128+1)/2 = 8256$
		bn4=$128(128+1)/2 = 8256$

Table 3.5: **(Continued from previous page)**

Model	**Extracted layers**	Gram matrix dimensionality

Inception-v3	`Conv2d_1a_3x3.bn`
	`Conv2d_2a_3x3.bn`
	`Conv2d_2b_3x3.bn`
	`Conv2d_3b_1x1.bn`
	`Conv2d_4a_3x3.bn`
		bn0=$32(32+1)/2 = 561$
		bn1=$32(32+1)/2 = 561$
		bn2=$64(64+1)/2 = 2080$
		bn3=$80(80+1)/2 = 3240$
		bn4=$192(192+1)/2 = 185208$

ResNet18	`layer1.0.bn1`
	`layer1.0.bn1`
	`layer2.0.bn1`
	`layer3.0.bn1`
	`layer4.0.bn1`
		bn0=$64(64+1)/2 = 2080$
		bn1=$64(64+1)/2 = 2080$
		bn2=$128(128+1)/2 = 8256$
		bn3=$256(256+1)/2 = 32896$
		bn4=$512(215+1)/2 = 131280$

ResNet34	`layer1.0.bn1`
	`layer1.0.bn1`
	`layer2.0.bn1`
	`layer3.0.bn1`
	`layer4.0.bn1`
		bn0=$64(64+1)/2 = 2080$
		bn1=$64(64+1)/2 = 2080$
		bn2=$128(128+1)/2 = 8256$
		bn3=$256(256+1)/2 = 32896$
		bn4=$512(215+1)/2 = 131280$

ResNet50	`layer1.0.bn1`
	`layer1.0.bn1`
	`layer2.0.bn1`
	`layer3.0.bn1`
	`layer4.0.bn1`
		bn0=$64(64+1)/2 = 2080$
		bn1=$64(64+1)/2 = 2080$
		bn2=$128(128+1)/2 = 8256$
		bn3=$256(256+1)/2 = 32896$
		bn4=$512(215+1)/2 = 131280$

ResNet101	`layer1.0.bn1`
	`layer1.0.bn1`
	`layer2.0.bn1`
	`layer3.0.bn1`
	`layer4.0.bn1`
		bn0=$64(64+1)/2 = 2080$
		bn1=$64(64+1)/2 = 2080$
		bn2=$128(128+1)/2 = 8256$
		bn3=$256(256+1)/2 = 32896$
		bn4=$512(215+1)/2 = 131280$

ResNet152	`layer1.0.bn1`
	`layer1.0.bn1`
	`layer2.0.bn1`
	`layer3.0.bn1`
	`layer4.0.bn1`
		bn0=$64(64+1)/2 = 2080$
		bn1=$64(64+1)/2 = 2080$
		bn2=$128(128+1)/2 = 8256$
		bn3=$256(256+1)/2 = 32896$
		bn4=$512(215+1)/2 = 131280$

VGG-16	`bn idx=[1, 8, 15, 25, 35]`	bn1=$64(64+1)/2 = 2080$ bn8=$128(128+1)/2 = 8256$ bn15=$256(256+1)/2 = 32896$ bn25=$512(215+1)/2 = 131280$ bn35=$512(215+1)/2 = 131280$
VGG-19	`bn idx=[1, 8, 15, 28, 41]`	bn1=$64(64+1)/2 = 2080$ bn8=$128(128+1)/2 = 8256$ bn15=$256(256+1)/2 = 32896$ bn25=$512(215+1)/2 = 131280$ bn35=$512(215+1)/2 = 131280$

Table 3.6: **Summary of the values from Brain-Score** benchmarks that we selected for our study (columns), for the 12 CNNs from our pool (rows). In the second column and second row, for models and benchmarks respectively, we report the .csv coordinates of the leaderboard. Our Brain-Score values were exported from the leaderboard website on September 8th 2025.

Model	Model coordinates	average vision	neural vision	behavior vision	V1	V2	V4	IT
Benchmark coordinates		D2	E2	DB2	F2	AM2	AP2	AV2
AlexNet	B290	0.171	0.281	0.06	0.338	0.176	0.355	0.255
Densenet-121	B128	0.311	**0.344**	0.278	**0.388**	**0.275**	**0.396**	0.315
DenseNet-169	B158	0.276	0.305	0.248	0.346	0.178	0.384	0.312
DenseNet-201	B94	0.318	0.293	0.343	0.34	0.19	0.388	0.253
Inception-v3	B84	0.33	0.3	0.358	0.321	0.169	**0.388**	0.324
Resnet18	B354	0.152	0.221	0.084	0.091	0.151	0.377	0.263
Resnet34	B357	0.147	0.206	0.089	0.077	0.143	0.359	0.244
Resnet50	B60	**0.362**	0.326	**0.397**	0.356	0.191	0.395	**0.363**
Resnet101	B176	0.26	0.081	0.44	0.127	0.068	0.062	0.066
Resnet152	B228	0.202	0.044	0.359	0.021	0.074	0.012	0.069
VGG-16	B104	0.322	0.288	0.354	0.319	0.147	0.385	0.303
VGG-19	B148	0.305	0.267	0.343	0.285	0.13	0.367	0.285

3.7.3 Supplementary figures

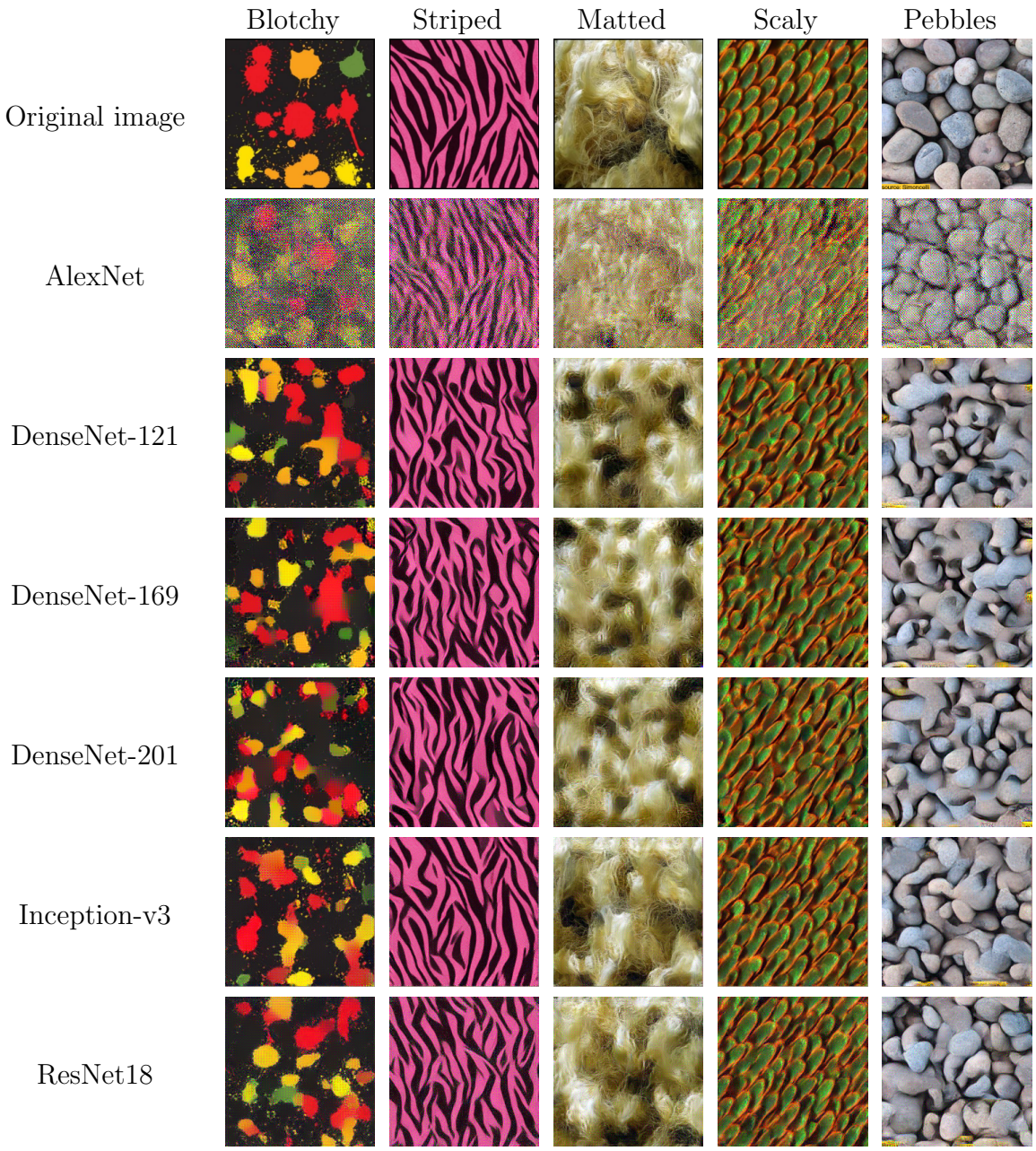


Table 3.7: **Textures generated with the Gatys algorithm** with the 12 CNNs of choice (rows). The 4 original images (top row) belong to 4 different classes in DTD (columns: *blotchy; striped; matted; scaly*). The last column represents one of the images originally used by Gatys et al., which we synthesized as a reference.

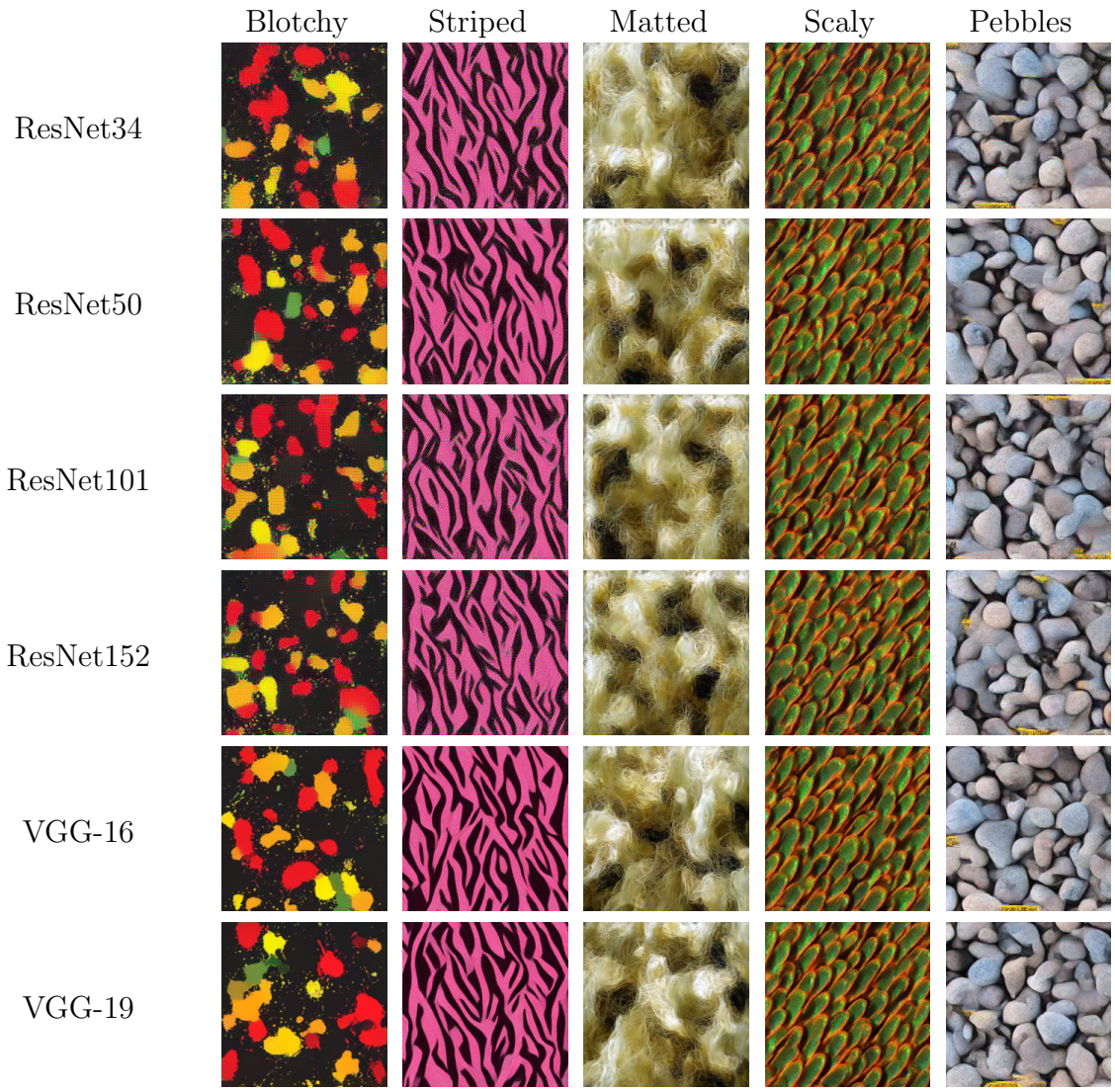


Table 3.8: **(Continued from previous page)**

Bibliography

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 929–947. ACM, 2024. doi: 10.1145/3620665.3640366.

Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019. doi: 10.1126/science.aav9436.

Guillaume Berger and Roland Memisevic. Incorporating long-range consistency in cnn-based texture generation. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

Nalini Bhushan, A. Ravishankar Rao, and Gerald L. Lohse. The Texture Lexicon: Understanding the Categorization of Visual Texture Terms and Their Relationship to Texture Images. *Cognitive Science, 21*, pages 219–246, 1997.

Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLOS Computational Biology*, 15(4):e1006897, 2019. doi: 10.1371/journal.pcbi.1006897.

Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib Majaj, and James J. DiCarlo. Deep neural networks rival the representation of primate it cortical neurons. *Proceedings of the National Academy of Sciences*, 111(23):8519–8524, 2014.

T. Caelli and B. Julesz. On perceptual analyzers underlying visual texture discrimination: Part i. *Biological Cybernetics*, 28(3):167–175, 1978. doi: 10.1007/BF00337138.

Riccardo Caramellino, Eugenio Piasini, Andrea Buccellato, Anna Carboncino, Vijay Balasubramanian, and Davide Zoccolan. Rat sensitivity to multipoint statistics is predicted by efficient coding of natural scenes. *eLife*, 10:e72081, 2021. doi: 10.7554/eLife.72081.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Ludovica de Paolis. Supplementary materials: Representational dissimilarity matrix plots for texture representations in cnns., April 2025. URL https://doi.org/10.5281/zenodo.19814745.

Jenelle Feather, Guillaume Leclerc, Aleksander Madry, and Josh H. McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034, 2023. doi: 10.1038/s41593-023-01442-0.

Jeremy Freeman, Corey M. Ziemba, David J. Heeger, Eero P. Simoncelli, and J. Anthony Movshon. A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, 16(7):974–981, 2013. doi: 10.1038/nn.3402.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. doi: 10.1523/JNEUROSCI.5023-14.2015.

Shouchang Guo, Valentin Deschaintre, Douglas Noll, and Arthur Roullier. U-attention to textures: Hierarchical hourglass vision transformer for universal texture synthesis. In *Proceedings of the 19th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–10. Association for Computing Machinery, 2022. doi: 10.1145/3565516.3565525.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Eric Heitz, Kenneth Vanhoey, Thomas Chambon, and Laurent Belcour. A sliced wasserstein loss for neural texture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9412–9420, 2021. doi: 10.1109/CVPR46437.2021.00929.

Margaret M. Henderson, Michael J. Tarr, and Leila Wehbe. A texture statistics encoding model reveals hierarchical feature selectivity across human visual cortex. *Journal of Neuroscience*, 43(22):4144–4161, 2023. doi: 10.1523/JNEUROSCI.1822-22.2023.

Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19000–19015. Curran Associates, Inc., 2020.

Ann M. Hermundstad, John J. Briguglio, Mary M. Conte, Jonathan D. Victor, Vijay Balasubramanian, and Gašper Tkačik. Variance predicts salience in central sensory processing. *eLife*, 3:e03722, 2014. doi: 10.7554/eLife.03722.

A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

G. Huang, Z. Liu, L. Van Der Maaten K. Q., and Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

Md. Amirul Islam, Matthew Kowal, Patrick Esser, Sen Jia, Björn Ommer, Konstantinos G. Derpanis, and Neil D. B. Bruce. Shape or texture: Understanding discriminative features in CNNs. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, volume 9906 of *Lecture Notes in Computer Science*, pages 694–711, Cham, 2016. Springer. ISBN 978-3-319-46475-6. doi: 10.1007/978-3-319-46475-6_43.

Bela Julesz. Visual pattern discrimination. *IRE Transactions on Information Theory*, 8: 84–92, 1962.

Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981. doi: 10.1038/290091a0.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PNAS*, 111(23):8619–8624, 2014.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. CORnet: Modeling the neural mechanisms of core object recognition. *bioRxiv*, page 408385, 2018. doi: 10.1101/408385.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Diversified texture synthesis with feed-forward networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3920–3928, July 2017.

Gang Liu, Yann Gousseau, and Gui-Song Xia. Texture synthesis through convolutional neural networks and spectrum constraints. In *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR)*, pages 3234–3239. IEEE, 2016. doi: 10.1109/ICPR.2016.7900133.

Li Liu, Jie Chen, Paul Fieguth, Guoying Zhao, Rama Chellappa, and Matti Pietikäinen. From bow to cnn: Two decades of texture representation for texture classification. *International Journal of Computer Vision*, 127(1):74–109, 2019. doi: 10.1007/s11263-018-1125-z.

David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2003. ISBN 9780521642989. Hardback, 640 pages, Published September 2003.

Jitendra Malik and Pietro Perona. Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A*, 7(5):923–932, 1990. doi: 10.1364/JOSAA.7.000923.

Simone Marsili. ndd, 2021. https://github.com/simomarsili/ndd.

Giorgio Matteucci, Eugenio Piasini, and Davide Zoccolan. Unsupervised learning of mid-level visual representations. *Current Opinion in Neurobiology*, 84:102834, 2024.

Paolo Muratore, Sina Tafazoli, Eugenio Piasini, Alessandro Laio, and Davide Zoccolan. Prune and distill: similar reformatting of image information along rat visual cortex and deep neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 30206–30218, 2022.

Paolo Muratore, Alireza Alemi, and Davide Zoccolan. Unraveling the complexity of rat object vision requires a full convolutional network and beyond. *Patterns*, 6(2):101149, 2025. Patterns (N Y).

Aran Nayebi, Nathan C. L. Kong, Chengxu Zhuang, Justin L. Gardner, Anthony M. Norcia, and Daniel L. K. Yamins. Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. *PLOS Computational Biology*, 19(10):e1011506, 2023. doi: 10.1371/journal.pcbi.1011506.

Ilya Nemenman, Fariel Shafee, and William Bialek. Entropy and inference, revisited. In *Advances in Neural Information Processing Systems 14*, pages 471–478, 2002.

Gouki Okazawa, Satohiro Tajima, and Hidehiko Komatsu. Gradual development of visual texture-selective properties between macaque areas v2 and v4. *Cerebral Cortex*, 27(10):4867–4880, 2017. doi: 10.1093/cercor/bhw282.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000. doi: 10.1023/A:1026553619983.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018. doi: 10.1523/JNEUROSCI.0388-18.2018.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018. URL https://www.biorxiv.org/content/10.1101/407007v2.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 2020. URL https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pages 1–14, 2015.

Katherine R. Storrs, Tim C. Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, 33(10):2044–2064, 2021. doi: 10.1162/jocn_a_01755.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

Tiberiu Tesileanu, Mary M. Conte, John J. Briguglio, Ann M. Hermundstad, Jonathan D. Victor, and Vijay Balasubramanian. Efficient coding of natural scene statistics predicts discrimination thresholds for grayscale textures. *eLife*, 9:e54347, 2020. doi: 10.7554/eLife.54347.

Tiberiu Tesileanu, Eugenio Piasini, and Vijay Balasubramanian. Efficient processing of natural scenes in visual cortex. *Frontiers in Cellular Neuroscience*, 16:1006703, 2022. ISSN 1662-5102. doi: 10.3389/fncel.2022.1006703.

Gašper Tkačik, Jason S. Prentice, Jonathan D. Victor, and Vijay Balasubramanian. Local statistics in natural scenes predict the saliency of synthetic textures. *Proceedings of the National Academy of Sciences*, 107(42):18149–18154, 2010. doi: 10.1073/pnas.0914916107.

TorchVision maintainers and contributors. TorchVision: PyTorch's Computer Vision library, November 2016. URL https://github.com/pytorch/vision.

M. R. Turner. Texture discrimination by gabor functions. *Biological Cybernetics*, 55(2):71–82, 1986. doi: 10.1007/BF00341922.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6924–6932, July 2017.

Jonathan Vacher, Aida Davila, Adam Kohn, and Ruben Coen-Cagli. Texture interpolation for probing visual perception. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22146–22157, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

J. D. Victor and M. M. Conte. Local image statistics: maximum entropy constructions and perceptual salience. *Optical Society of America*, 29:1313–1345, 2012.

Jonathan D. Victor, Mary M. Conte, and Charles F. Chubb. Textures as probes of visual processing. *Annual Review of Vision Science*, 3:275–296, 2017.

Kasper Vinken, Xavier Boix, and Gabriel Kreiman. Incorporating intrinsic suppression in deep neural networks captures dynamics of adaptation in neurophysiology and perception. *Science Advances*, 6(42):eabd4205, 2020. doi: 10.1126/sciadv.abd4205.

Yifan Wang, Aleksander Holynski, Brian L. Curless, and Steven M. Seitz. Infinite texture: Text-guided high-resolution diffusion texture synthesis. *arXiv preprint arXiv:2405.08210*, 2024. doi: 10.48550/arXiv.2405.08210.

Felix A. Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 9:501–524, 2023. doi: 10.1146/annurev-vision-120522-031739.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23): 8619–8624, 2014.

Ning Yu, Connelly Barnes, Eli Shechtman, Sohrab Amirghodsi, and Michal Lukac. Texture mixer: A network for controllable synthesis and interpolation of texture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12164–12173, June 2019.

Yunguo Yu, Anita M. Schmid, and Jonathan D. Victor. Visual processing of informative multipoint correlations arises primarily in v2. *eLife*, 4:e06604, 2015. doi: 10.7554/eLife.06604.

Mirko Zanon, Bastien S. Lemaire, Eugenio Piasini, Riccardo Caramellino, Caroline Nallet, Vijay Balasubramanian, Judit Gervain, Davide Zoccolan, and Giorgio Vallortigara. Predisposed and learned preferences for multipoint visual statistics in visually naïve newly hatched chicks. *bioRxiv*, 2025. bioRxiv preprint.

Kalathupiriyan A. Zhivago and Sripati P. Arun. Texture discriminability in monkey inferotemporal cortex predicts human texture perception. *Journal of Neurophysiology*, 112(11):2745–2755, 2014. doi: 10.1152/jn.00532.2014.

S. C. Zhu, X. W. Liu, and Y. N. Wu. Exploring texture ensembles by efficient markov chain monte carlo: Toward a "trichromacy" theory of texture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):554–569, 2000. doi: 10.1109/34.862195.

Corey M. Ziemba, Jeremy Freeman, J. Anthony Movshon, and Eero P. Simoncelli. Selectivity and tolerance for visual texture in macaque v2. *Proceedings of the National Academy of Sciences*, 113(22), 2016.

Corey M. Ziemba, Robbe L. T. Goris, Gabriel M. Stine, Richard K. Perez, Eero P. Simoncelli, and J. Anthony Movshon. Neuronal and behavioral responses to naturalistic texture images in macaque monkeys. *Journal of Neuroscience*, 44(42), 2024.

CHAPTER 4

Texture Representations in Deep Vision Models: Comparing CNNs, Vision Transformers, and Human Perception

This chapter is a reproduction of the paper Texture Representations in Deep Vision Models: Comparing CNNs, Vision Transformers, and Human Perception, currently in *arXiv* as an e-print.
Statement of contribution: I performed all experiments and analyses, made all the figures and wrote the first draft of the paper. Alessandro Laio provided assistance with the implementation of Information Imbalance and with optimizing the plots. Eugenio Piasini provided the code for the implementation of one texture synthesis algorithm. Marco Baroni, Alessandro Laio and Eugenio Piasini provided scientific guidance, contributed to the conceptual framing and discussion of the findings.

4.1 Abstract

In computational vision science, Convolutional Neural Networks (CNNs) have emerged as a popular model of biological vision because of the alignment they can exhibit with neural and behavioral data in humans and animals. However, it remains unclear to what extent this alignment persists for visual tasks that extend beyond the canonical object-recognition paradigm based on well-defined semantic content. In this study, we diverge from the common object-centric view by focusing on another aspect of vision: texture perception. We consider textures of different complexity generated with three different algorithms from the same source images. Using a rank-based statistic, we quantify the information encoded in the internal representations of a CNN and three Vision Transformers (ViTs), and we compare the similarity of these representations to those inferred from human psychophysics data. We find that the representation of textures is aligned in different ViTs, but not between the ViTs and the CNN; that ViTs form similar representations for textures of different complexity; that human performance in recognizing textures can be better predicted from ViTs representations rather than CNN representations. Taken together, these results suggest that ViTs may capture more faithfully than CNNs how texture patterns are visually processed by humans, and that the representations of texture stimuli in computational models may be driven by the network architecture.

4.2 Introduction

Convolutional neural networks (CNNs) are ubiquitous in computer vision and in vision neuroscience, where they are commonly employed as models of the ventral visual stream and of the biological mechanisms underlying object recognition (Kriegeskorte, 2015; Lindsay, 2021). A pivotal moment in the history of CNNs was the release of AlexNet (Krizhevsky et al., 2012), which demonstrated significantly higher accuracy

than other vision models in an object classification task connected to the ImageNet database (Deng et al., 2009). This success reinforced the centrality of the object recognition paradigm in machine learning and computational vision neuroscience, which led to numerous advances in the following years. However, in real life scenarios image perception and representation is much more diversified than just recognizing objects. Consequently, parallel research directions in vision science have been focusing on aspects of vision that cannot be addressed within the framework of object recognition. One of these tasks is texture perception: visual textures are an acknowledged method to study complex real world vision phenomena (Victor et al., 2017). An early formal definition of textures was provided by Julesz (Julesz, 1962, 1981), stating that textures can be seen as families of patterns that share certain local regularities. Julesz also conjectured that textures could be modeled via low-order statistics (Julesz et al., 1973). This conjecture became very influential in the development of computational models of textures, which later resulted in the release of texture synthesis algorithms that operationalized the conjecture by means of low-level correlations. The most prominent models are those by Portilla & Simoncelli (Portilla and Simoncelli, 2000a), Victor & Conte (Victor and Conte, 2012), and Gatys et al., (Gatys et al., 2015). The model by Victor & Conte synthesizes artificial binary textures by means of local correlations in 2×2 pixel gliders in a 10-dimensional parametrized space. The model by Portilla & Simoncelli produces gray-scale textures by means of a fixed set of statistics of the wavelet coefficients of a given input image. The algorithm by Gatys synthesizes naturalistic textures by computing local correlations among the feature maps extracted from a CNN pretrained on object recognition. Since CNNs are regarded as a good model for object recognition in human and primate vision (Güçlü and van Gerven, 2015; Lindsay, 2021; Schrimpf et al., 2018; Storrs et al., 2021; Yamins et al., 2014), the success of this approach indirectly linked texture perception mechanisms to those underlying object recognition. However, it was recently shown that the quality of the texture representation associated with different CNNs does not correlate with well-established metrics of representational and behavioral alignment of these same networks with biological vision (de Paolis et al., 2026). This finding suggests that common measures of alignment may focus too narrowly on object recognition, and may underweight other perceptual phenomena like texture perception. This, in turn, may imply that network architectures different from CNNs may be useful representational models when more diverse tasks other than object classification are taken into account.

In the last few years, Vision Transformers (ViT) have emerged as another extremely effective architecture for computer vision tasks (Chen et al., 2020). ViT can perform object classification (Radford et al., 2021), feature distillation (Oquab et al., 2024) and image generation (Chen et al., 2020; Yu et al., 2022). ViTs belong to the family of Transformers, which are deep neural network models equipped with the attention mechanism operationalized with multi-head attention layers, which allow to produce internal representations of the input images encoded in an array of tokens by attending all the possible combinations of positions between image patches of a fixed size. While CNNs produce image representation leveraging local correlations with convolutions and pooling operations, ViTs form image representations by means of both local and long range correlations among image patches. The attention mechanism effectively integrates both levels of correlations, which in turn produce image representations robust to spatially-bound information, allowing ViTs to form stable and consistent representations across data that preserve the same meaning or structure.

We are interested in studying how CNNs and ViTs represent textures, in comparing their performance, and in comparing the models' performance against human subjects'

perception. We perform a comparative analysis of CNNs, Vision Transformers, and human subjects using a large image dataset spanning a continuum of visual complexity. The dataset combines two existing stimulus databases, one containing images of objects and one containing images of textures. Starting from the texture database, we generate synthetic textures with three different algorithms, producing stimuli with distinct levels of complexity. We then extract texture representations from one CNN and three ViTs and compare them with human behavioral performance in a texture-perception task. The models' representations are analyzed using a rank-based statistic designed to quantify information in high-dimensional feature spaces. Human performance is measured with an odd-one-out paradigm, as in Jagadeesh and Gardner (2021); Wallis et al. (2017), in which participants identify the odd stimulus among sets of similar textures with different levels of complexity.

We find that (i) the three ViTs align among themselves in representing textures of different types of complexity, but the CNN does not align with ViTs; (ii) human perception of textures of different types strongly aligns with the texture representations produced by the three ViTs but does not align with the texture representations produced by the CNN.

4.3 Methods

4.3.1 Stimulus datasets

Textures and objects databases. To create the dataset for our study we assemble two existing databases, one containing images of objects and one of textures. We use the Describable Textures Dataset (DTD) as a texture dataset and as a source to generate synthetic textures with the three algorithms by Gatys et al. (2015); Portilla and Simoncelli (2000a); Victor and Conte (2012). Using this procedure, we obtain a large pool of textures with distinct levels of complexity, where each datapoint is generated from one datapoint of the source, DTD. In this way, our natural and synthesized texture datasets are naturally aligned at the single-image level. This passage is key as the rank-based metric of our choice quantifies the variations in information across different subsets as changes in the local neighborhoods.

As shown in Figure 4.1, the textures generated with the Victor & Conte algorithm (**V&C**) are simple, coarse, and show little structure; the textures by Portilla & Simoncelli (**P&S**) are intermediate, greyscale and preserve some global structure; the textures by the Gatys' algorithm (**G**) are more complex, show good local structure and color; finally, the **DTD** textures are the most perceptually and semantically complex. We expand the description of the synthetic textures below. DTD was developed by Cimpoi et al. (2014) to have both naturalistic quality and human-like semantic accuracy. It consists of 5640 images of realistic textures collected from the internet, equally distributed in 47 semantic categories. The class labels are selected from the "Texture Lexicon", an independent study by Bhushan et al. (1997) in which participants were asked to build a set of words describing categories of textures. The naturalistic status of its textures and the independent linguistic human agreement make DTD relevant for our scope, because the textures contain relevant semantic information.

For the objects database, we considered a subset of 5640 images from ImageNet-1K validation set (Deng et al., 2009; Russakovsky et al., 2015) (henceforth **Object**). ImageNet is the most popular dataset for training neural networks in computer vision and the most used for studying object classification and perception with computational models. We therefore use it as a reference stimulus for representing complex objects with

a well-defined semantic content provided by the class labels. Finally, we also consider 5640 images constituted by fully random black and white pixels (henceforth **Noise**) to exemplify images with neither semantic content nor lower-level systematic structure.

Synthetic texture stimuli. As mentioned above, we use three different synthesis algorithms: **V&C** by Victor & Conte, **P&S** by Portilla & Simoncelli and **G** by Gatys, to generate synthetic stimuli that show different levels of image complexity. We use textures from DTD as an input for each algorithm as follows:

- **V&C:** We generated V&C type of textures using the implementation by Piasini (2021) of Victor & Conte's model. This method samples maximum-entropy textures with fixed correlation statistics among neighboring pixels. The algorithm provides an elegant parameterization of texture space, which is suitable for highly-controlled experiments, but it produces textures that are perceptually distant from those found in natural images: they feature black and white squares arranged in a regular pattern. These textures contain no semantic content, and constitute the lowest-level subset in our semantic content hierarchy. (Figure 4.1: V&C).

- **P&S:** We generated P&S type textures using the Python adaptation of the original Matlab implementation by Portilla & Simoncelli, publicly available in the GitHub repository textureSynth by Portilla and Simoncelli (2000b). This algorithm computes the correlations among parameters of a wavelet decomposition, generating synthetic textures that statistically match the ones found in input images. The generated textures are in grayscale, preserve most of the structure in periodic images, and serve as image samples of mid-level complexity (Figure 4.1: P&S).

- **G algorithm:** We generated G type textures with the algorithm by Gatys et al. (2015), implemented by de Paolis et al. (2026) publicly available in the GitHub repository perceptual_misalignment. The algorithm effectively defines a texture ensemble based on the Gram matrices formed by computing spatial correlations among intermediate layer activations of a CNN trained for object classification (VGG-19, Simonyan and Zisserman (2015)). Texture synthesis is achieved by optimizing the pixels of an initially random image to match the statistical structure of the original texture via gradient descent. The generated textures preserve the colors and the perceptual content of the original image, as well as most of the semantic content, serving as image samples with middle-to-high complexity (Figure 4.1: G). Further details on the algorithms' implementation and data preprocessing are provided in Appendix 4.7.3.

4.3.2 Models

We extracted the features of all the stimulus datasets described above from all the layers of four vision neural networks, in order to understand how they represent different kinds of textures and run a comparative analysis. We tested three ViTs and one CNN. The models vary in architecture, training objective and task, but they share the training dataset: object images from ImageNet (Deng et al., 2009). For the ViT family, we tested the visual encoder of **CLIP** (Radford et al., 2021), a supervised ViT pretrained with a contrastive image-text objective; **DINO-v2** (Oquab et al., 2024), a self-supervised image features extractor; **iGPT** (Chen et al., 2020), an unsupervised autoregressive decoder for

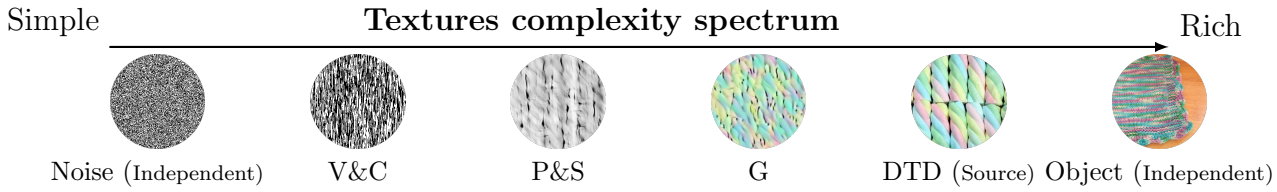


Figure 4.1: **Example stimuli arranged along a putative spectrum of texture complexity.** In the Noise sample, each pixel is sampled independently at random as black or white with equal probability. The stimuli labeled as V&C, P&S, and G are obtained by the DTD image classified as "braided" (marked as "source") and by application of the respective algorithm. The Object image was chosen by hand for illustrative purposes from the Object dataset.

image generation. For the CNN family we tested **VGG-19** (Simonyan and Zisserman, 2015), a supervised image classifier commonly used as a model of image processing in the ventral visual stream (Geirhos et al. (2017); Schrimpf et al. (2018)). When performing this analysis, we controlled for parameter count, task, training objective, and architecture. VGG-19, CLIP, DINO-v2 number of parameters is around the same order of magnitude; VGG-19 and CLIP can perform object classification; all the models were trained with different objectives; but ViTs share the attention mechanism in their architecture, while VGG-19 has convolutions. A summary of the models' details is provided in Appendix 4.7.4, Table 4.1.

4.3.3 Human psychophysics

We want to obtain a measure of human agreement in forming relationships between different types of textures, and to compare their performance to that of CLIP, DINO-v2, iGPT, and VGG-19 respectively. We designed a psychophysics experiment to measure the perceptual distance between textures belonging to different subsets as described above, by assessing the similarity between texture images belonging to the same semantic class ("braided") but displaying different complexity levels. Inspired by the psychophysics literature on texture perception, we adopted an odd-one-out paradigm (Jagadeesh and Gardner, 2021; Wallis et al., 2017), implemented as follows:

- on each trial, three image stimuli are presented simultaneously, at three different locations that are equidistant from a fixation cross in the center of a computer screen;
- two of them (the "reference" stimuli) belong to the same subset, while the third (the "odd") to a different one (e.g. two DTD vs. one P&S);
- the participants' task is to identify the odd stimulus;
- we only consider texture stimuli from V&C, P&S, G and DTD subsets;
- within the same trial, the three stimuli belong to the same DTD semantic class;
- stimuli are shown for 200ms, preventing visual search and engaging pre-attentive judgment.
- following a standard procedure (**?**), we displayed all stimuli in grayscale regardless of their subset or class.

We collected data from 51 participants. Further details about the experimental setup, a visual display of the stimuli, and an illustration of the experimental flow can be found in Appendix 4.7.5, Figures 4.10, 4.11.

4.3.4 Information Imbalance

We use Information Imbalance (**II**) (Glielmo et al., 2022b) as a metric to compare the representations of all models with respect to each subset. II is a non-symmetric measure of similarity that relies on the number of shared neighbors for each datapoint of two spaces A and B, by quantifying how much information is shared between the two as the similarity between local neighborhoods. Since it relies on local distances, II is robust with respect to representational spaces lying on high-dimensional manifolds (Acevedo et al., 2026b). II is estimated as:

$$2/N^2 \sum_{i,j} r_{ij}^A = 1 r_{ij}^B \tag{4.1}$$

where A and B are the two spaces, r^A and r^B are the neighborhood ranks of one point for A and one point for B, N is the total number of datapoints. If II is closer to 1, the two spaces are not mutually predictable; if it is around 0.5, they are mutually and symmetrically predictable; if it is closer to 0, one space predicts the other. II is an asymmetric measure of predictability when it is closer to 1 or 0, but since our results were mostly symmetric with II values around 0.5, we adopted a symmetrized value of II, by averaging the two resulting values from any analysis (Acevedo et al., 2026a). For this study we used an adaptation of the implementation of II from the public repository DADApy (Glielmo et al., 2022a).

4.4 Results

We extracted the features of all the datasets (see Figure 4.1) and from all the models. For VGG-19 we extracted the batch-normalization layer following each convolutional block, consistently with previous literature (de Paolis et al., 2026; Gatys et al., 2015). For CLIP, DINO-v2, and iGPT we considered the average of each layer's final hidden state as a global representation of the token sequence; in the case of CLIP and DINO-v2 we excluded the end-of-sequence CLS tokens. Since ViT activations show the presence of massive outliers connected to the attention sinking effect (Kang et al., 2025; Kaul et al., 2025; Son et al., 2024), we clipped the activations of CLIP, DINO-v2, and iGPT to retain the .95 percentile (Kim et al., 2021; Ma et al., 2019). This passage was unnecessary for VGG-19 activations, as they tend to be more stable.

We then performed the following analyses: (i) a comparative analysis of the models' representation for each subset, to assess whether the models generate similar or dissimilar representations for the same type of image; (ii) a comparative analysis across the representations of the image types generated from the same DTD source – therefore Noise, V&C, P&S, G, plus DTD – for each model, to assess whether each model can form representations that display meaningful relationships among the images; (iii) we compared the models' response to the previous analysis against human performance collected in the psychophysics experiment, to test whether the representations across textures found in the ViTs align to human perception better than those of the CNN; (iv) the study of the effect of different images on the representational geometry of all models (see Appendix 4.7.1).

ViTs show agreement in representing the same image type. Here we compare the models' representations for each type of input image. For each dataset, we compute II among all the possible pairs of models. In figure 4.2, we show a graph illustration of the results of this analysis: for each subset, one graph has four nodes, one per model, and six links, indicating the $1 - II$ value computed for each pair of models. The links are unidirectional because we symmetrized II, and for illustrative purposes, we converted II into $1 - II$. We focus on the layer's representations with the largest $1 - II$ values for each model pair. A full plot for all the layers can be found in Appendix 4.7.2.

When Noise is the input dataset, we observe only thin links connecting the models, with extremely low $1 - II$ values: this happens because the representations formed by the four models are extremely different and there is no mutual predictability between them. If we observe the case of V&C, we notice a radical change: already when representing simple textures, there are strong links among CLIP, DINO-v2, and iGPT, as well as weak links between VGG-19 and the three ViTs. The strong links mean that the representations formed by CLIP, DINO-v2, and iGPT are similar and predictable by each other, while the thin links between VGG-19 and the three ViTs signify non-mutually predictable representations. This pattern keeps being present and evolving as the textures become semantically richer: in the cases of P&S and G, CLIP, DINO-v2 and iGPT still hold strong links among each other, while links with VGG-19 are weak. When DTD is the input dataset, the situation changes: links among ViTs are still strong, but the links to VGG-19 become stronger, with $1 - II \simeq 0.5$. This means that the representations between VGG-19 and the ViTs become mutually predictable. In the case of Object, we still see strong links among ViTs, and increasingly stronger links involving VGG-19, with $1 - II \simeq 0.75$: in particular, the Object graph shows that the representations of complex and semantically rich images are largely shared and aligned across all models, regardlessly of their architecture.

These results suggest that VGG-19 and ViTs rely on different representational strategies. While ViTs capture the representations of the same textures as early as in V&C, and maintain this representational pattern stable through the complexity evolution of textures until DTD, VGG-19 instead seems to form representations of textures that differ according to the complexity: semantically poor textures elicit representations that do not align well to those of the ViTs, while semantically rich textures representations and objects representations largely align to those of the ViTs.

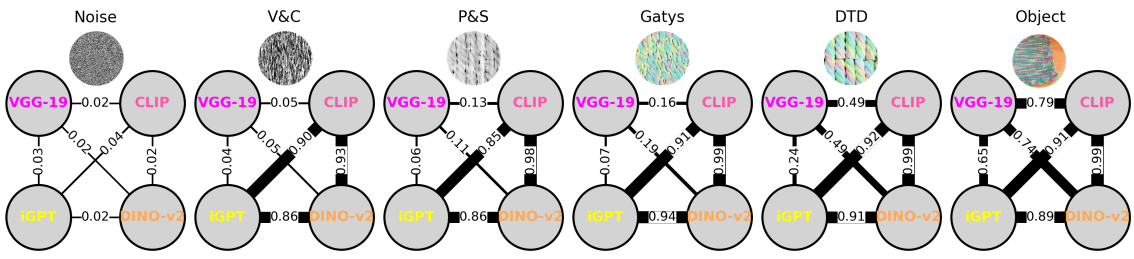


Figure 4.2: **Graphs representing the symmetrized mean II across models.** For visualization purposes, for each subset we selected the best II between the two most mutually predictable layers for all models. For ease of illustration, we used $1 - II$: strong links and high values show high mutual predictability; thin links and low values represent low predictability. An extended figure is available in Appendix 4.7.2, Figure 4.7.

ViTs capture relationships among textures of different types. In this analysis, we investigate the relationship between different textures in each model's representation. For each model, we compute II between all the possible combinations of textures: we consider Noise, V&C, P&S, G, and DTD. Once again we show a graph representation of the results of this analysis in figure 4.3: for each model, one graph has five nodes, one for each type of input image, and ten links, one for each pair between input types. Also in this case the links are unidirectional because we symmetrized II values. For illustrative purposes, we converted II to $1-II$, and we visualized only the model's layer representation with the best II values. A full plot of the analysis with all the layers can be found in Appendix 4.7.2.

The graph of VGG-19 shows weak links connecting any data type considered, Noise, V&C, P&S, G, DTD, where the best $1-II$ is around ~ 0.1. This means that the representations of VGG-19 of any image type are unrelated to each other. When we consider CLIP instead, we observe strong links and $(1-II) \geq 0.9$ among P&S, DTD, and G textures, as well as medium sized links with $(1-II) \sim 0.35$ with respect to V&C textures. We instead observe thin links with $(1-II) \sim 0$ between Noise and the four types of textures. This result shows that CLIP representations of P&S, DTD and G are very similar and highly mutually predictable, V&C representations are somewhat mutually predictable to the other three textures, and Noise representations are not mutually predictable with respect to any texture. Both DINO-v2 and iGPT show the same pattern as CLIP: strong links among P&S, DTD, and G, medium-strenght links between V&C and the other three textures, extremely weak links between Noise and all the textures.

These results suggest that while CLIP, DINO-v2, and iGPT form largely overlapping representations between high and middle-level textures as in P&S, DTD, G, and form quite similar representations to V&C, VGG-19, instead, does not produce shared or overlapping representations among any image type. The absence of any representational similarity with respect to Noise in any of the three models is not surprising, since it belongs to a different distribution other than that of the DTD source. In this analysis, we excluded Object because a point-wise comparison with the other subset using II is unfeasible.

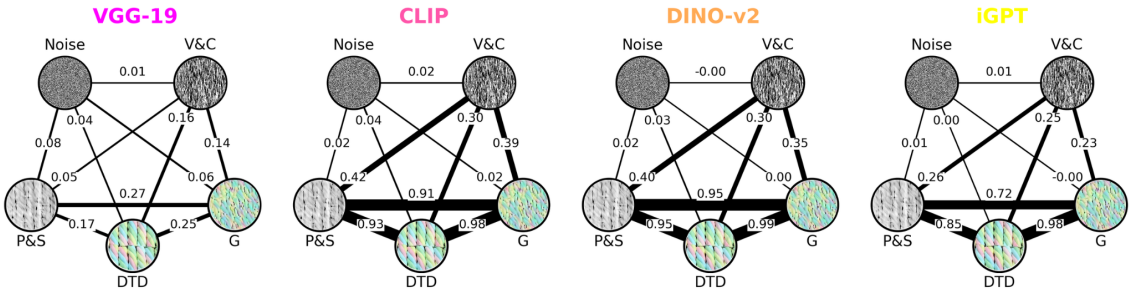


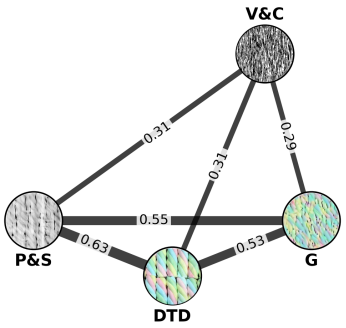
Figure 4.3: **Graphs representing the symmetrized mean II across datasets**, excluding Object. For representational purposes, for each model we selected the II between the best mutually predictive layers of all the subset combinations. For illustrative purposes, we inverted the II scale to $1-II$: strong links correspond to high mutual predictability; thin links represent low predictability; slightly thick links represent equal mutual predictability. An extended figure is available in Appendix 4.7.2, Figure 4.8.

ViTs align with human perception in representing textures. Finally, we compare the similarity across datasets according to the various models, discussed in the previous analysis, against human performance, to test which model best aligns with human

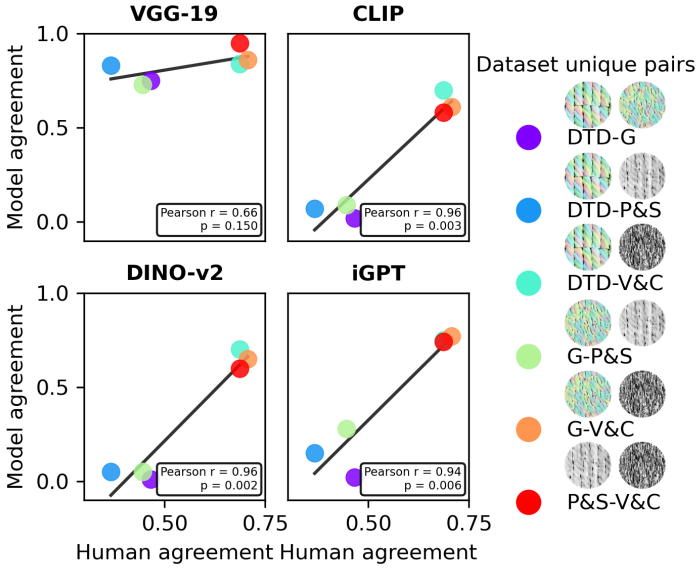
perception of textures.

We obtained one human accuracy score for each pair of texture subsets by averaging accuracy across participants, and we represented accuracy values as undirected links in a graph. Finally, we run a correlation test with Pearson's r to assess which model best matches human perception. For visualization purposes, figure 4.4a shows a graph representation of human $1 - accuracy$ scores as links among V&C, P&S, G, and DTD textures, represented as nodes. We observe thick links and medium-high $1 - accuracy$ scores among P&S, DTD, and G, plus medium-strength links V&C and the three other textures. This result shows that participants easily distinguish V&C textures from the rest of the textures, when V&C is presented as the odd; on the other hand, humans find more difficult to tell apart P&S, DTD, and G when either of them is presented as an odd. This indicates that high-to-medium level complexity textures are perceived as similar to each other, while semantically poor textures are perceived as outliers, that are thus easier to tell apart. Figure 4.4b shows the result of the correlation analysis. The colored dots are II values found in the analysis in paragraph 4.4, coded in the legend for each pair of textures, that are correlated to the participants' accuracy scores. We observe that, while VGG-19 does not show any significant correlation with human performance, CLIP, DINO-v2 and iGPT all show high and significant correlation values.

These results suggest that ViTs representations of textures that display different levels of complexity align better to human texture perception than those of VGG-19.



(a) Human accuracy graph.



(b) Correlation between human accuracy and model II scores.

Figure 4.4: **Human pairwise accuracy and its correlations with models performance.** For illustrative purposes, in panel (a) we represent $1 -$ accuracy, so that thicker links mean stronger perceptual overlap. Panel (b) shows the four correlation plots among humans and the four models. We used Pearson's r and we report r, p inside the boxes. The legend maps a color for the II score used in the correlation for each pair of image types.

4.5 Discussion

In this study we investigated the representation of visual textures in artificial neural networks and the alignment between deep neural network representations and human perception of these textures. We built a large dataset of visual textures characterized by different degrees of perceptual complexity. We considered two existing databases: ImageNet for object images, and Describable Textures Dataset for naturalistic textures; we then used three texture synthesis algorithms to generate textures with progressively simpler content starting from DTD: Gatys' algorithm, Portilla & Simoncelli's algorithm, Victor & Conte's model. We also considered a set of randomly generated images.

We explored four models: VGG-19, CLIP, DINO-v2 and iGPT. VGG-19 is a CNN and a popular model of the mammalian ventral visual stream; CLIP, DINO-v2, and iGPT are ViTs, and their alignment to human vision is largely unexplored. We extracted features from all models with all the image types: Noise, V&C, P&S, G, DTD, Object. We collected human accuracy scores with a odd-one-out texture discrimination task and analyzed the relationship among the models' representations in several ways with Information Imbalance, as follows: (i) to quantify the agreement between the representations of the same image type for all models; (ii) to quantify the relationship captured by the models' representations for different textures; (iii) to compare human perception against the models' performance.

We find that ViTs produce overlapping representation of all textures and Object, while VGG-19's representations are different and become similar to those of ViTs only when DTD and Object serve as input. We therefore conclude that the architecture may be the main factor explaining the difference in performance between the ViTs and the CNN as it an important shared element among the three ViTs.

We furthermore observe that while the three ViTs can form largely overlapping representations of P&S, DTD, and G, and somewhat similar representations between those textures and V&C, VGG-19 doesn't form any similar representation among the four types of textures. Once again, we connect this result to the architecture: ViTs are equipped with attention, designed to build representations with both local and global range correlations among tokens (Chen et al., 2020; Vaswani et al., 2017). Convolutions and pooling instead generate representations that rely only on spatial-bound and local image features, although effective spatial ranges grow with layer depth. We conclude that attention plays a role in the formation of image representations that are stable and shared across models with the same architecture but different objectives, consistently with the Platonic Hypothesis (Huh et al., 2024), which are robust to the progressive perceptual loss of the stimuli.

Finally, we find that human perception of texture is better aligned with the ViT representations than with the CNN representations. This result suggests that ViTs, despite not being commonly recognized as a good model of human vision, outperform VGG-19 in at least certain perceptual dimensions.

Taken together, our results suggest that ViTs might be useful models of human texture perception, and should be considered as valid models to expand research about human visual perception, perceptual organization and biological modeling. More broadly, our findings underscore the idea that texture and object processing are only partially intertwined. Building robust models of texture perception may therefore require moving beyond systems trained exclusively for object recognition, while still accounting for the contribution of high-level semantic information. Our results also point to a potential limitation in the way Julesz's conjecture has been operationalized in computational models

of texture perception. Rather than relying exclusively on local image feature correlations, our findings suggest that such models may need to adopt a broader representational perspective than that originally implied by Julesz's conjecture.

4.6 Limitations

Due to limited computational resources, we could not include other CNN models in the analyses. If this could be done, based on the representational similarities for textures across different CNN architectures (de Paolis et al., 2026), we would expect them to show similar results as VGG-19 and to be dissimilar to ViTs.

Bibliography

Santiago Acevedo, Alessandro Laio, and Marco Baroni. Differential syntactic and semantic encoding in LLMs, 2026a.

Santiago Acevedo, Andrea Mascaretti, Riccardo Rende, Matéo Mahaut, Marco Baroni, and Alessandro Laio. A quantitative analysis of semantic information in deep representations of text and images, 2026b. URL https://arxiv.org/abs/2505.17101.

Nalini Bhushan, A. Ravishankar Rao, and Gerald L. Lohse. The Texture Lexicon: Understanding the Categorization of Visual Texture Terms and Their Relationship to Texture Images. *Cognitive Science, 21*, pages 219–246, 1997.

Riccardo Caramellino, Eugenio Piasini, Andrea Buccellato, Anna Carboncino, Vijay Balasubramanian, and Davide Zoccolan. Rat sensitivity to multipoint statistics is predicted by efficient coding of natural scenes. *eLife*, 10:e72081, 2021. doi: 10.7554/eLife.72081.

Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Ludovica de Paolis, Fabio Anselmi, Alessio Ansuini, and Eugenio Piasini. Perceptual misalignment of texture representations in convolutional neural networks. *arXiv preprint arXiv:2604.01341*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Robert Geirhos, David Janssen, Heiko H. Schütt, Jonas Rauber, Matthias Bethge, and Felix A. Wichmann. Of human observers and deep neural networks: A detailed psychophysical comparison. *Journal of Vision*, 17(10):806, 2017. doi: 10.1167/17.10.806.

Aldo Glielmo, Iuri Macocco, Diego Doimo, Matteo Carli, Claudio Zeni, Romina Wild, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Dadapy: Distance-based analysis of data-manifolds in python. *Patterns*, page 100589, 2022a. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2022.100589. URL https://www.sciencedirect.com/science/article/pii/S2666389922002070.

Aldo Glielmo, Claudio Zeni, Bingqing Cheng, Gabor Csanyi, and Alessandro Laio. Ranking the information content of distance measures. *PNAS Nexus*, 1(2):pgac039, 2022b. doi: 10.1093/pnasnexus/pgac039.

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. doi: 10.1523/JNEUROSCI.5023-14.2015.

Ann M. Hermundstad, John J. Briguglio, Mary M. Conte, Jonathan D. Victor, Vijay Balasubramanian, and Gašper Tkačik. Variance predicts salience in central sensory processing. *eLife*, 3:e03722, 2014. doi: 10.7554/eLife.03722.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Akshay V. Jagadeesh and Justin L. Gardner. Texture-like representation of objects in human visual cortex. *Nature Neuroscience*, 24:224–231, 2021. doi: 10.1038/s41593-020-00751-2.

Bela Julesz. Visual pattern discrimination. *IRE Transactions on Information Theory*, 8: 84–92, 1962.

Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981. doi: 10.1038/290091a0.

Bela Julesz, E. N. Gilbert, L. A. Shepp, and H. L. Frisch. Inability of humans to discriminate between visual textures that agree in second-order statistics revisited. *Perception*, 2(4):391–405, 1973.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/da8a39bc39ae1c89dd6ebb1e3bcbb3f3-Paper-Conference.pdf.

Prannay Kaul, Chengcheng Ma, Ismail Elezi, and Jiankang Deng. From attention to activation: Unraveling the enigmas of large language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/556c2ff2ca01492f14f52d079bdc55f3-Paper-Conference.pdf.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. Self-guided contrastive learning for BERT sentence representations. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.197. URL https://aclanthology.org/2021.acl-long.197/.

Nikolaus Kriegeskorte. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446, 2015. doi: 10.1146/annurev-vision-082114-035447.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc., 2012.

Grace W. Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, 33(10):2017–2031, 2021. doi: 10.1162/jocn_a_01544.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. Universal text representation from bert: An empirical study. *arXiv preprint arXiv:1910.07973*, 2019. doi: 10.48550/arXiv.1910.07973.

Matéo Mahaut and Marco Baroni. Similarity of Processing Steps in Vision Model Representations. 2026. doi: 10.48550/arXiv.2601.21621.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193. arXiv:2304.07193.

Eugenio Piasini. Metex - maximum entropy textures, v1.2.0, 2021. URL https://zenodo.org/records/18500242.

Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000a. doi: 10.1023/A:1026553619983.

Javier Portilla and Eero P. Simoncelli. textureSynth: Matlab code for analyzing and synthesizing visual textures, 2000b. https://github.com/LabForComputationalVision/textureSynth.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018. URL https://www.biorxiv.org/content/10.1101/407007v2.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pages 1–14, 2015.

Seungwoo Son, Wonpyo Park, Woohyun Han, Kyuyeun Kim, and Jaeho Lee. Prefixing attention sinks can mitigate activation outliers for large language model quantization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-main.134/.

Katherine R. Storrs, Tim C. Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, 33(10):2044–2064, 2021. doi: 10.1162/jocn_a_01755.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

J. D. Victor and M. M. Conte. Local image statistics: maximum entropy constructions and perceptual salience. *Optical Society of America*, 29:1313–1345, 2012.

Jonathan D. Victor, Mary M. Conte, and Charles F. Chubb. Textures as probes of visual processing. *Annual Review of Vision Science*, 3:275–296, 2017.

Thomas S. A. Wallis, Christima M. Funke, Alexander S. Ecker, Leon A. Gatys, Felix A. Wichmann, and Mathias Bethge. A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of Vision*, 17: 1–29, 2017.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23): 8619–8624, 2014.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In *International Conference on Learning Representations*, 2022.

Mirko Zanon, Bastien S. Lemaire, Eugenio Piasini, Riccardo Caramellino, Caroline Nallet, Vijay Balasubramanian, Judit Gervain, Davide Zoccolan, and Giorgio Vallortigara. Predisposed and learned preferences for multipoint visual statistics in visually naïve newly hatched chicks. *bioRxiv*, 2025. bioRxiv preprint.

4.7 Appendix

4.7.1 Supplementary analysis

Effect of texture type on each model's signature geometry.

We focus on the effect of the type of input image on the representational geometry pattern of each model. For each model and each data type, we compute the symmetrized II between all the pair combinations of layers' representations. The matrices in figure 4.5 show the representational geometry of the models induced by the data type. Each matrix color coding is relative to the specific data-per-model combination and own II values, represented in logarithmic scale and reported in the legend at the side. The relative maximum II value is colored in yellow, while the minimum value, always at 0, is colored in dark blue.

The first row reports the internal representations of VGG-19. If we observe the matrix for VGG-19/Objects, the entries are the II values for each possible combinations of layers (for details about the number of layers see Appendix Table 4.1). In this case, the input stimuli elicit a representational pattern featuring two dark clusters, one at the top of the diagonal and the other at the bottom. The information shared among close layers is relatively high because their representations have II values close to ~ 0. The color of the heatmap goes progressively towards yellow as we move away from the diagonal, because II values are around the maximum. In this case, the pairwise layer representations do not share much information. This happens for layer pairs that are distant from each other. In the case VGG-19/DTD, we observe a similar color pattern to that of VGG-19/Objects: two darker clusters, one at the top of the diagonal and one at the bottom, where representation of close layers are similar; when moving away from the diagonal yellow prevails because the representations between distant layers become dissimilar. This representational pattern seems to fade in VGG-19/G and VGG-19/P&S, and it gets completely disrupted in VGG-19/V&C: here, in fact, we observe a large dark block at the bottom half of the diagonal, where representations between the last group of layers are similar, but also redundant because they don't display further information shifts. The representations between the other layers' pair combinations feature lighter colors, signaled by intermediate II values due to a certain degree of similarity.

Moving on to the second row and observing the heatmap of CLIP/Object, we notice two main dark blocks, one located at the top half of the diagonal and the other at the bottom half. A dark coloration is present around the diagonal, which becomes progressively lighter as the entries are II values calculated between distant layers' representations. This pattern is consistent to what we observed in VGG-19/Objects: Objects elicit representations that tend to be similar between closer layers, with II values around ~ 0, and dissimilar between layers located farther away, with II values towards the maximum. In the case of CLIP/DTD, we observe a similar color pattern, with less defined clusters, but still featuring lower II values around the diagonal. The pattern starts to break from DTD/G and across DTD/P&S, where we observe an uniform dark coloration

stemming from the diagonal, which only gets lighter when layer n.1 is compared against all the other layers. This means that CLIP representations of respectively G and P&S textures are similar across the whole model, with a single information shift between the input layer and the rest of the layers. The pattern completely disappears in CLIP/V&C: the information shared between layers n.1 and n.2 is quite low, while it is redundantly high in the representations between rest of the layers' combination. The representational patterns observed in CLIP show a similar progression to those of VGG-19, but while the difference in information through layers in CLIP appears to evolve smoothly, in VGG-19 we observe occasional breaks signaled by lighter stripes: we connect it to the effect of the pooling on the beginning of each of the 5 convolutional blocks.

DINO-v2's representations follow a similar evolution to that observed in CLIP: DINO-v2/Objects shows dark blocks around the diagonal, with a large cluster in the upper half of the diagonal and a second cluster in the bottom half. The blue central part, featuring low II values among close layers, fades to yellow when II is computed among distant layers. We find the same pattern in DINO-v2/DTD, with less defined clusters around the diagonal. This characteristic representational geometry, similarly to CLIP, gradually dissipates as the textures become simpler, through G, P&S, and V&C: in these cases, the clusters become less and less structured, resulting in a general darkening of the whole heatmap, except dark stripes in the input layer where II is ~ 0.

In the row of iGPT, we find a slightly different signature geometry, probably due to the following characteristics: it organizes its representations of images as an encoder-decoder even if the architecture backbone is an decoder only (Chen et al., 2020; Mahaut and Baroni, 2026; Valeriani et al., 2023); it is deeper than the other architectures (see Appendix Table 4.1); the objective is next-pixel prediction. In iGPT/Object we observe an almost uniform medium-dark coloration through all the layers, with four yellow spots in the middle of the four sides of the matrix: this means that the representation is generally self-similar among all layers, however they become most dissimilar when II is computed between intermediate layers and early layers, and between intermediate layers and late layers, respectively. In practice, iGPT representational shift happens around the bottleneck at the center, while in VGG-19, CLIP, and DINO-v2 the representational bottleneck is around the output layer, as we show with more classical linear methods (see Appendix Supplementary Figure 4.6). Also in the case of iGPT the signature pattern progressively dissipates as the input textures becomes simpler: in iGPT/DTD the representational geometry resembles that of iGPT/Object, with a meaningful information shift in the middle layers. Then, through iGPT/G and iGPT/P&S, the information shift in the middle becomes gradually less confined to a few middle layers. And finally, in iGPT/V&C, II tends to be redundantly low, the middle-layers information bottleneck disappears, switching to a single, initial information shift between early layers and the rest.

In the case of Noise images, across VGG-19, CLIP, DINO-v2, and iGPT layers representations, we always observe a prominent yellow color given by high II values, accompanying a darker area surrounding the diagonal. The only exception is iGPT, where dark and light colors with respectively low and high II values, both occupy about half of the heatmap. In all models, the representational geometry elicited by Noise is compatible to representations that are always changing and share little information with previous layers. This is due to the randomness and unpredictability of the information carried in noisy images, which at any processing step elicit a certain degree of surprisal in the models.

Taken together, these results suggest that the representations of Objects and those of semantically rich texture such as DTD rely on a similar strategy in all models: the

representations in each group of close layers are well structured. However, as the textures gradually lose semantic content and become simpler, the representational strategy adapts to the lack of semantic content of the images and becomes more redundant with only one relevant shift: most information is elaborated in the first layers, while the rest is just repetition. Surprisingly, while the four textures types belong to the same distribution of the source database DTD, only the representations of DTD seem to resemble those of Object, even though Object is drawn from a different distribution with a different semantic labeling system. We conclude that the type of input, and in particular its semantic content, has an effect on the representational geometry of all models. Semantically rich images and textures show representations structured as clusters, while semantically poor textures produce more sparse representations. This experiments highlights that, despite differences in architecture and objectives between the CNN and the ViTs, the progressive loss of complexity in the input images has an effect on the representational organization of each model, which tends to rely on all layers in semantically-rich images, while it tends to rely only on the first layers in low-complexity textures.

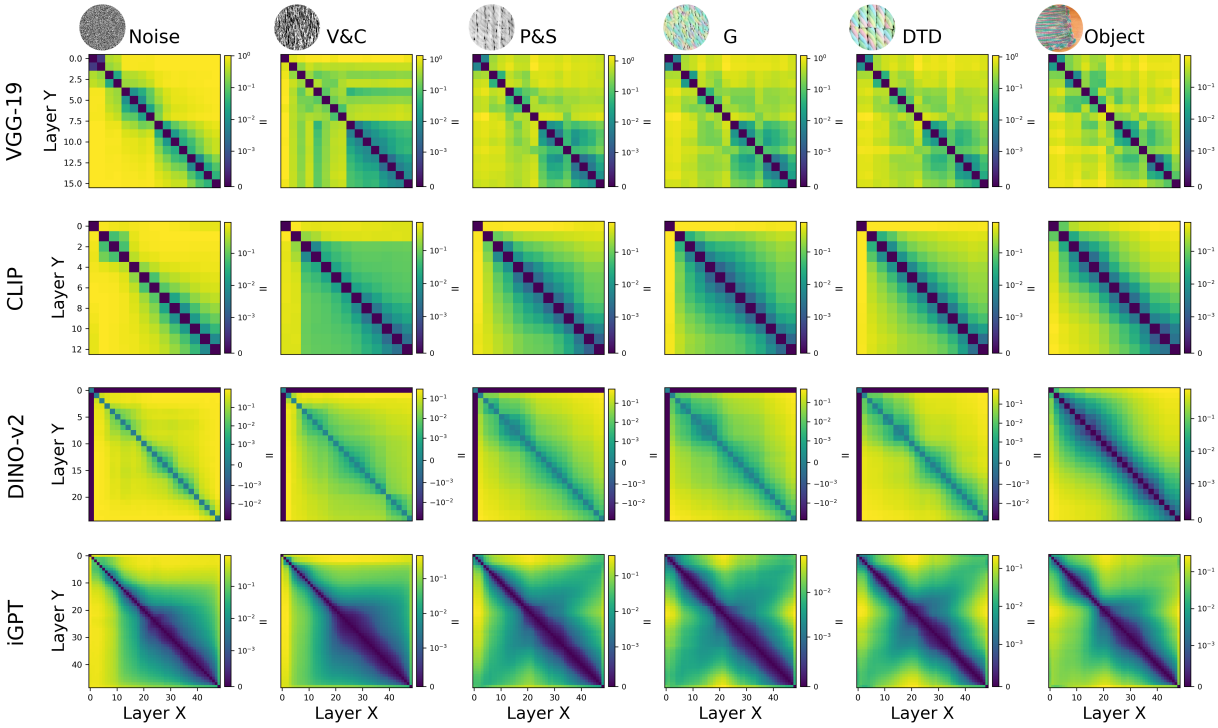


Figure 4.5: **Heatmaps for each model (rows) by each dataset (columns).** For each matrix, the entries are the symmetrized mean II values calculated between all the possible combinations of layers. Each matrix is equipped with its relative color-bar of II values in logarithmic scale.

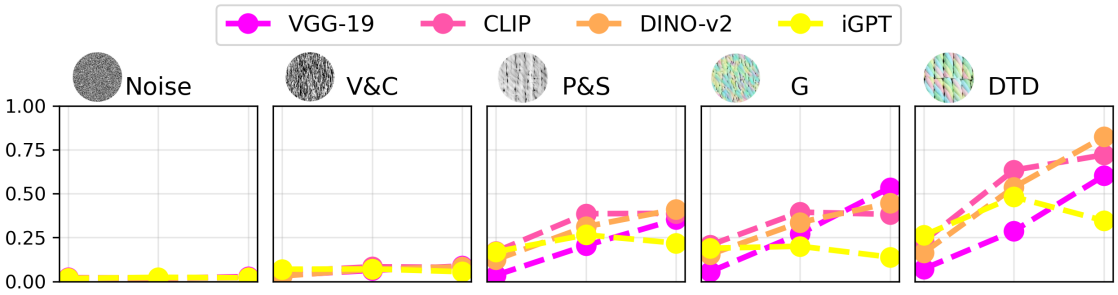
A linear probe does not transfer knowledge across models.

In order to assess whether the analysis done with Information Imbalance, a highly non-linear metric, can be also performed with linear model, we build a linear classifier to classify images of textures generated from the source DTD into the original labels found in DTD.

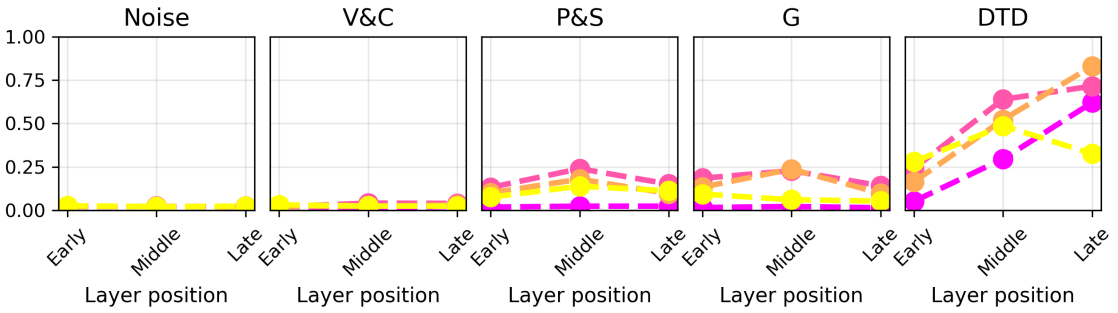
For each model we build three linear classifiers, one for an early, one middle, and one for a deep layer. We then perform the following experiments (a) classification: for

each model, we trained and tested three linear probes to classify their features into their original classes with a 80-20 train-test regime; (b) transfer learning: for each model, we trained the three probes to classify 80% of the DTD features and tested them to classify 20% of the features from the remaining subsets. We excluded Object subset because it does not share the classification structure with the other textures subsets.

Figure 4.6 shows that (a) the classifiers increase in accuracy as the data become more complex. iGPT peaks accuracy in the middle layer because it is an autoencoder, while the other models peak in the last layer. This behavior agrees with previous studies (Chen et al., 2020; Mahaut and Baroni, 2026; Valeriani et al., 2023); (b) the classifiers fail at generalizing features learned from DTD. We conclude that a linear classifier can't capture the relationship among subsets, while II can do so due to its nonlinearity.



(a) Classification



(b) Transfer learning

Figure 4.6: **Linear probe experiments. (a) Classification**: accuracy scores for a triplet of probes trained on early, middle, and deep extracted features for each dataset. **(b) Transfer learning**: accuracy scores for a triplet of probes trained on early, middle, and deep layers' extracted features of DTD and tested on the other datasets.

4.7.2 Supplementary figures

ViTs show agreement in representing the same image type.



Figure 4.7: **Extended plot for Section 4.2, Figure 4.2**: For each dataset (columns) we compute II for each pair of models. In particular, as shown in the rows, each model serves as a *reference*, and the other three serve as *target*, as shown in the columns and in the top legend. We quantify the depth of each feature's layer as the relative percentage, and compared symmetrized II scores between reference layers and target layers along the same relative depth.

ViTs capture relationships among textures of different types.

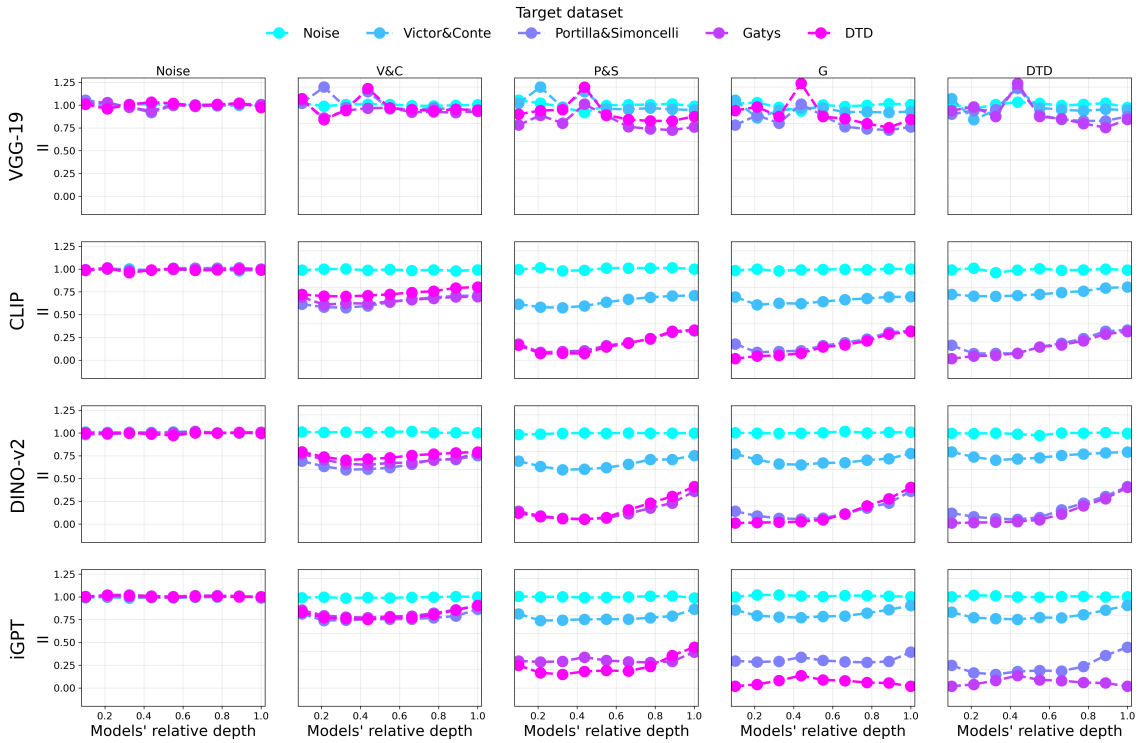


Figure 4.8: **Extended plot for Section 4.3, Figure 4.3**: For each model (rows) we compute II between all pairs combination between a *reference dataset* (columns) and four remaining *target datasets* (legend). We quantify the depth of each feature's layer as the relative percentage, and compare symmetrized II scores between reference layers and target layers along the same relative depth.

Human accuracy

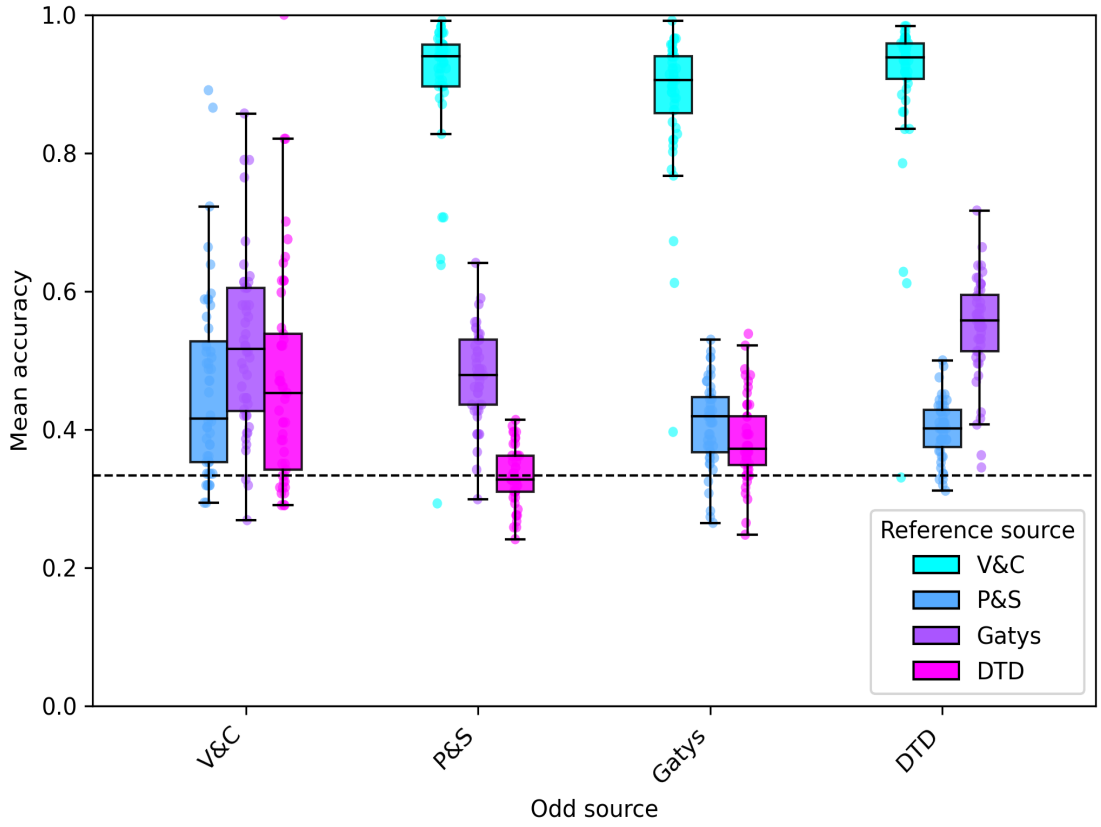


Figure 4.9: **Extended plot for Section 4.5, Figure 4.4a**: This plot represents the relationships among texture subsets for each pair of *odd image* (see X axis) and and *reference image* (legend) as human accuracy scores (see Y axis) of all participants collected in the behavioral task. Each box summarizes the distribution of mean accuracies for one combination of odd source and reference source. The scatter shows the individual data points, the box shows the interquantile range, the central line shows the median, and the errorbars show the non-outlier range. We observe that accuracy is highly odd source-dependent. Trials with V&C as the source were solved with almost top accuracy for all odd-source conditions, indicating that V&C images form a perceptually coherent and distinct subset. In contrast, when V&C was the odd source, performance was moderate: accuracy in distinguishing P&S, G and DTD gets lower, suggesting a substantial perceptual overlap. This asymmetry suggests that odd-one-out performance is not determined by a symmetric perceptual distance between datasets, but also by the internal variability of the reference source.

4.7.3 Stimulus dataset generation

Victor and Conte

We used this algorithm to generate textures of type **V&C**. This approach produces low-complexity binary textures that have been employed in a number of efficient coding studies (Caramellino et al., 2021; Hermundstad et al., 2014; Zanon et al., 2025) due to their appealing analytical properties and controllability and despite being perceptually distant

from natural stimuli. The algorithm allows to control explicitly all local image statistics measured on 2×2 pixel blocks (Victor and Conte, 2012). By exploiting an assumption of translational symmetry, the 16 possible block statistics are reparameterized into ten coordinates:

- luminance bias: $\{\gamma\}$
- pairwise correlations: $\{\beta_{_}, \beta_{|}, \beta_{\backslash}, \beta_{/}\}$
- three-point correlations: $\{\theta_{\llcorner}, \theta_{\lrcorner}, \theta_{\ulcorner}, \theta_{\urcorner}\}$
- four-point correlations: $\{\alpha\}$

Given chosen values of one or more of these coordinates, the remaining image statistics are fixed implicitly by a maximum-entropy principle: the generated texture is the most random binary image ensemble compatible with the specified local constraints. To generate the texture of type **V&C** corresponding to a given DTD image we first computed the block statistics following a procedure analogous to that described in Hermundstad et al. (2014). Then, we determined which among the 10 parameters was most strongly biased away from 0, the point in parameter space that indicates full randomness. Finally, we sampled a random texture constrained to have that same value of the parameter using METex (Piasini, 2021).

Portilla and Simoncelli

We used this algorithm to generate textures of type **P&S**. It works particularly well on regular and homogeneous textures, returning spatially regular and gray-scaled synthetic images of textures. Given an exemplar texture, the image is first decomposed using a steerable pyramid. The algorithm then measures a fixed set of statistics on the resulting wavelet coefficients. A synthetic texture is generated by starting from a random image sampled from $\hat{x} \sim \mathcal{N}(0, 1)$, and iteratively projecting the image onto the constraint sets defined by the target statistics, until the synthetic image matches the exemplar statistics.

We used the algorithm with the following parameters (see Portilla and Simoncelli (2000a) and Portilla and Simoncelli (2000b) for reference implementation of the algorithm):

- `-n`: depth of steerable pyramid = 5
- `-k`: number of orientations of steerable pyramid = 4
- `-n`: pixel distance for calculation of auto-correlations = 7
- `-iter`: number of iterations = 10

Gatys

We used this algorithm to generate textures of type **G**. Type G Textures are characterized by strong perceptual and semantic similarity with the original textures, resulting in high quality RGB images. The algorithm's backbone is VGG-19_bn, and the texture representations are extracted from 5 batch normalization layers compiled in to Gram matrices. For each of the 5 blocks in VGG-19 we extracted the first batch normalization layers after the first convolution of the block, resulting in 5 layers. Given an image from DTD, x, the

algorithm feeds x as input to VGG-19. For each extracted layer l, the feature activations F^l are used to compute a Gram matrix G^l, defined as:

$$G^l_{ij} = \sum_m F^l_{im} F^l_{jm}, \tag{4.2}$$

where the index i and j are the indices of the feature maps, and the index m runs over all units within a feature map. This Gram matrix summarizes the correlations between feature channels, discarding spatial information. Texture synthesis is performed by iteratively updating the pixels of an initially random image $\hat{x}$, where pixel values are drawn from $\hat{x} \sim \mathcal{N}(0, 1)$, such that its Gram matrices match those of the original image x. This is achieved by minimizing the *Gram loss*, defined as a weighted sum of L_2 distances between the Gram matrices of the individual layers:

$$\mathcal{L}(\hat{x}, x) = \sum_l \frac{1}{4M_l N_l^2} \|\hat{G}^l - G^l\|_2^2. \tag{4.3}$$

where the weighting factor is chosen to balance the contribution to the loss of each layer (de Paolis et al., 2026). To minimize the Gram loss we used Adam at 30000 optimization steps. This choice allowed us to control the quality of the generated texture, by preserving some of the semantic and perceptual content of the original textures in DTD without resulting too similar.

4.7.4 Models details

Model	Param.	Architecture	Objective	Extr. layers
VGG-19_bn	148M	CNN	Classification	16
CLIP-ViTB-32	302M	ViT	Image-text pairing	12
DINO-v2	304M	ViT	Feature extr.	24
iGPT	1,413M	ViT	Next token pred.	49

Table 4.1: **Summary of model's details.** Consistently with previous studies, we used VGG-19_bn. Indeed, batch normalization layers allow a stable feature representations of textures (de Paolis et al., 2026; Gatys et al., 2015). We extracted every batch normalization layer after any convolutional layer, for a total of 16 layers.

4.7.5 Human behavioral experiment

Experimental design

We recruited 51 participants on the platform Prolific, paid £11/hour. We did not apply any particular screening with respect to age, spoken language, country of residence. Participants were required to have a normal or corrected-to-normal vision and no past history of epilepsy. The stimuli were taken from our curated datasets, cropped in circles and converted to grayscale Wallis et al. (2017). The experiment consists in 47 blocks, one for each semantic class, composed of 30 trials, for a total of 1410 trials and a median time of 45 minutes. Images in each stimulation belong to the same semantic class. Participants were instructed to press these keys to indicate the odd: → for right image; ← for left image; ↓ for bottom image.

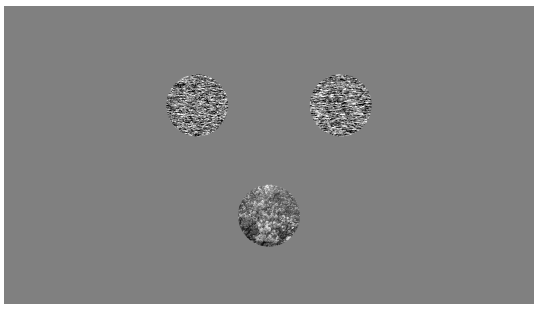


Figure 4.10: **Visual stimulus screen in the experimental design.** The odd in this case is the bottom stimulus (↓).

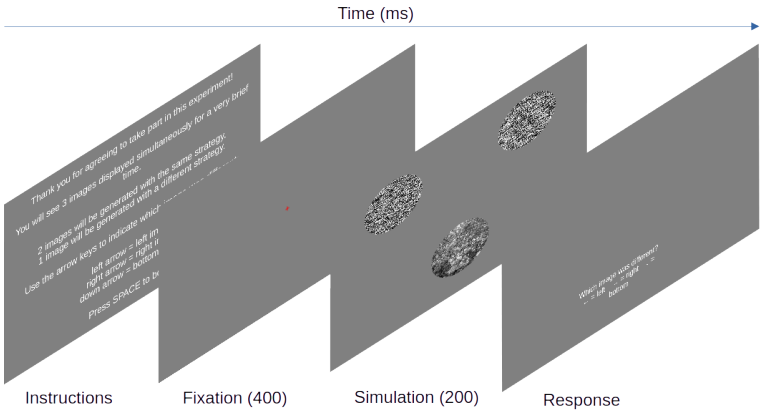


Figure 4.11: **Illustration of the experimental flow**, with time in ms. Each simulation is paced at 200ms, proceeded by a 400ms red fixation cross. Response time is self paced. Breaks between trials are provided.

Humans-models correlations scores

Model	Pearson's r	p-value
VGG-19_bn	0.66	0.150
CLIP	0.96	0.003
DINO-v2	0.96	0.002
iGPT	0.94	0.006

Table 4.2: **Summary of table humans-models correlations for Section 4.5, Figure 4.4b**.

Bibliography

Santiago Acevedo, Alessandro Laio, and Marco Baroni. Differential syntactic and semantic encoding in LLMs, 2026a.

Santiago Acevedo, Andrea Mascaretti, Riccardo Rende, Matéo Mahaut, Marco Baroni, and Alessandro Laio. A quantitative analysis of semantic information in deep representations of text and images, 2026b. URL https://arxiv.org/abs/2505.17101.

Nalini Bhushan, A. Ravishankar Rao, and Gerald L. Lohse. The Texture Lexicon: Understanding the Categorization of Visual Texture Terms and Their Relationship to Texture Images. *Cognitive Science, 21*, pages 219–246, 1997.

Riccardo Caramellino, Eugenio Piasini, Andrea Buccellato, Anna Carboncino, Vijay Balasubramanian, and Davide Zoccolan. Rat sensitivity to multipoint statistics is predicted by efficient coding of natural scenes. *eLife*, 10:e72081, 2021. doi: 10.7554/eLife.72081.

Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Ludovica de Paolis, Fabio Anselmi, Alessio Ansuini, and Eugenio Piasini. Perceptual misalignment of texture representations in convolutional neural networks. *arXiv preprint arXiv:2604.01341*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Robert Geirhos, David Janssen, Heiko H. Schütt, Jonas Rauber, Matthias Bethge, and Felix A. Wichmann. Of human observers and deep neural networks: A detailed psychophysical comparison. *Journal of Vision*, 17(10):806, 2017. doi: 10.1167/17.10.806.

Aldo Glielmo, Iuri Macocco, Diego Doimo, Matteo Carli, Claudio Zeni, Romina Wild, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Dadapy: Distance-based analysis of data-manifolds in python. *Patterns*, page 100589, 2022a. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2022.100589. URL https://www.sciencedirect.com/science/article/pii/S2666389922002070.

Aldo Glielmo, Claudio Zeni, Bingqing Cheng, Gabor Csanyi, and Alessandro Laio. Ranking the information content of distance measures. *PNAS Nexus*, 1(2):pgac039, 2022b. doi: 10.1093/pnasnexus/pgac039.

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. doi: 10.1523/JNEUROSCI.5023-14.2015.

Ann M. Hermundstad, John J. Briguglio, Mary M. Conte, Jonathan D. Victor, Vijay Balasubramanian, and Gašper Tkačik. Variance predicts salience in central sensory processing. *eLife*, 3:e03722, 2014. doi: 10.7554/eLife.03722.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Akshay V. Jagadeesh and Justin L. Gardner. Texture-like representation of objects in human visual cortex. *Nature Neuroscience*, 24:224–231, 2021. doi: 10.1038/s41593-020-00751-2.

Bela Julesz. Visual pattern discrimination. *IRE Transactions on Information Theory*, 8: 84–92, 1962.

Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981. doi: 10.1038/290091a0.

Bela Julesz, E. N. Gilbert, L. A. Shepp, and H. L. Frisch. Inability of humans to discriminate between visual textures that agree in second-order statistics revisited. *Perception*, 2(4):391–405, 1973.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/da8a39bc39ae1c89dd6ebb1e3bcbb3f3-Paper-Conference.pdf.

Prannay Kaul, Chengcheng Ma, Ismail Elezi, and Jiankang Deng. From attention to activation: Unraveling the enigmas of large language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/556c2ff2ca01492f14f52d079bdc55f3-Paper-Conference.pdf.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. Self-guided contrastive learning for BERT sentence representations. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.197. URL https://aclanthology.org/2021.acl-long.197/.

Nikolaus Kriegeskorte. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446, 2015. doi: 10.1146/annurev-vision-082114-035447.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc., 2012.

Grace W. Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, 33(10):2017–2031, 2021. doi: 10.1162/jocn_a_01544.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. Universal text representation from bert: An empirical study. *arXiv preprint arXiv:1910.07973*, 2019. doi: 10.48550/arXiv.1910.07973.

Matéo Mahaut and Marco Baroni. Similarity of Processing Steps in Vision Model Representations. 2026. doi: 10.48550/arXiv.2601.21621.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193. arXiv:2304.07193.

Eugenio Piasini. Metex - maximum entropy textures, v1.2.0, 2021. URL https://zenodo.org/records/18500242.

Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000a. doi: 10.1023/A:1026553619983.

Javier Portilla and Eero P. Simoncelli. textureSynth: Matlab code for analyzing and synthesizing visual textures, 2000b. https://github.com/LabForComputationalVision/textureSynth.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018. URL https://www.biorxiv.org/content/10.1101/407007v2.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pages 1–14, 2015.

Seungwoo Son, Wonpyo Park, Woohyun Han, Kyuyeun Kim, and Jaeho Lee. Prefixing attention sinks can mitigate activation outliers for large language model quantization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-main.134/.

Katherine R. Storrs, Tim C. Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, 33(10):2044–2064, 2021. doi: 10.1162/jocn_a_01755.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

J. D. Victor and M. M. Conte. Local image statistics: maximum entropy constructions and perceptual salience. *Optical Society of America*, 29:1313–1345, 2012.

Jonathan D. Victor, Mary M. Conte, and Charles F. Chubb. Textures as probes of visual processing. *Annual Review of Vision Science*, 3:275–296, 2017.

Thomas S. A. Wallis, Christima M. Funke, Alexander S. Ecker, Leon A. Gatys, Felix A. Wichmann, and Mathias Bethge. A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of Vision*, 17: 1–29, 2017.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23): 8619–8624, 2014.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In *International Conference on Learning Representations*, 2022.

Mirko Zanon, Bastien S. Lemaire, Eugenio Piasini, Riccardo Caramellino, Caroline Nallet, Vijay Balasubramanian, Judit Gervain, Davide Zoccolan, and Giorgio Vallortigara. Predisposed and learned preferences for multipoint visual statistics in visually naïve newly hatched chicks. *bioRxiv*, 2025. bioRxiv preprint.

CHAPTER 5
Conclusions

This thesis investigated visual perception through the study of textures, within the broader goal of understanding what information is represented by computational models of vision and how these representations relate to human perception. Rather than treating vision primarily as a problem of object recognition, I focused on visual textures as stimuli that occupy an intermediate position between simple image statistics and semantically meaningful natural scenes. Across the three projects presented in this thesis, the central question progressively shifted from how to generate and parameterize naturalistic textures to how their information is internally represented, and ultimately to whether the representations formed by deep vision models resemble those underlying human perception.

Chapter 1 approached this problem from a generative perspective. I attempted to combine the Gram matrix-based texture representation introduced by Gatys et al. (2015) with the latent space of a Variational Autoencoder. The objective was to obtain a continuous and interpretable representational manifold of naturalistic textures that could eventually be explored in relation to human perception. The complete model did not achieve this goal. Although our implementation of the Gatys algorithm for texture synthesis successfully synthesized naturalistic textures and the deterministic autoencoder could reconstruct increasingly complex images, introducing the variational bottleneck substantially degraded performance. The results indicated a tension between compressing images into a Gaussian latent distribution and preserving the high-dimensional, multiscale statistics constrained by Gram matrices. This negative result was nevertheless important for the development of the thesis: by separating and testing the components of the model, it identified the Gram representation as the most robust element of the framework and motivated the subsequent projects.

Chapter 2 focused directly on Gram matrix-based representations. I expanded my implementation of the Gatys representation across twelve CNNs trained for object recognition and quantified how well their internal texture representations recovered the human-defined categorical structure of the Describable Textures Dataset (Cimpoi et al., 2014). Texture information generally increased with network depth, showing that deeper CNN representations retain progressively more information relevant to perceptual texture categories. However, even the best representations recovered only part of the available class structure. More importantly, their quality showed no significant relationship with Brain-Score measurements (Schrimpf et al., 2018) of neural or behavioral alignment with the mammalian visual system. Thus, CNNs that constitute better models of biological vision according to standard benchmarks, are not necessarily better models of texture representation.

Chapter 3 addressed the limitation highlighted in the previous study by including Vision Transformers and, crucially, direct measurements of human perception. I constructed a visual stimulus space ranging from simple synthetic textures to naturalistic textures and objects and compared their representations in VGG-19 (Simonyan and Zisserman, 2015), CLIP (Radford et al., 2021), DINO-v2 (Oquab et al., 2024), and iGPT (Chen et al., 2020) using Information Imbalance Glielmo et al. (2022b). Despite their different training objectives, the three Vision Transformers produced remarkably consistent relationships among texture types. These relationships were preserved across substantial changes in texture complexity, whereas VGG-19 representations varied much more strongly across the same stimuli. Human psychophysics provided an independent benchmark: participants per-

ceived intermediate and naturalistic textures as strongly related, while simpler textures were progressively more distinguishable. This perceptual organization was significantly correlated with the representations of all three Vision Transformers, but not with those of VGG-19.

Overall, this thesis shows that texture perception probes aspects of visual representation that cannot be reduced straightforwardly to object recognition. A model that successfully predicts object-related neural activity or behavior may organize texture information differently from humans, and, conversely, architectures not traditionally regarded as biological models of vision can provide a better account of particular perceptual dimensions. These observations supports a broader conception of computational vision modeling in which models are evaluated across multiple visual functions rather than against a single dominant task. Visual textures provide a particularly useful ground to test this approach because their statistical structure can be systematically manipulated while their relationship to natural perception is preserved. Extending this framework to a wider range of architectures, neural measurements, and psychophysics could therefore help characterize which computational principles are genuinely shared between artificial and biological vision.