Texture and art with deep neural networks

Leon A. Gatys,[1,2,3] Alexander S. Ecker,[1,2,5] Matthias Bethge,[1,2,4]

[1]Werner Reichardt Centre for Integrative Neuroscience and Institute of Theoretical Physics, University of Tübingen, Germany
[2]Bernstein Center for Computational Neuroscience, Tübingen, Germany
[3]Graduate School for Neural Information Processing, Tübingen, Germany
[4]Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[5]Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA

Abstract

Although the study of biological vision and computer vision attempt to understand powerful visual information processing from different angles, they have a long history of informing each other. Recent advances in texture synthesis that were motivated by visual neuroscience have led to a substantial advance in image synthesis and manipulation in computer vision using Convolutional Neural Networks (CNNs). Here we review these recent advances and discuss how they can in turn inspire new research in visual perception and computational neuroscience.

Highlights

- State-of-the-art convolutional neural networks extract perceptual image properties
- CNNs allow to edit and synthesise high-level perceptual variables in images
- CNNs hold great potential for creating digital art
- Computational Neuroscience should try to explain CNN function

Introduction

A fascinating property of human visual perception is that physically very different images are perceived to look very much the same. A prominent example of this property is texture perception: When more than a handful of similar objects are nearby, our visual system groups them together and we become insensitive to their precise spatial arrangement – 'things' become 'stuff' (Fig. 1A) (Adelson and Bergen, 1991; Dubuc and Zucker, 2001)

Since texture perception is omnipresent in human vision, it has occupied vision scientists for many years to characterise under what conditions things become stuff and what exactly constitutes a texture. Mathematically, we can formalise a texture as a sample from an ensemble of images with (spatially) stationary statistics. That is, the local statistical dependencies between pixels are the same irrespective of absolute position and, consequently, individual elements ("Textons", Julesz, 1984) can have a highly variable spatial arrangement (Fig 1B).

The study of visual textures as stationary images was pioneered by Julesz (Julesz, 1962), who hypothesised that all images with equal N^{th}-order joint pixel histograms are pre-attentively indistinguishable for human observers and therefore samples from the same texture. This specific hypothesis turned out to be wrong for computationally tractable values of N (Julesz et al., 1978, 1973). Yet, the basic idea to describe a texture by a set of spatial summary statistics forms the basis of parametric texture modelling and even the notion of images with equal N^{th}-order joint pixel histograms is still applied fruitfully today (Purpura et al., 1994; Yu et al., 2015).

The summary statistics that define the texture do not have to be defined at the level of image pixels, but can employ other feature representations. A number of researchers explored feature spaces that resemble the response properties of the early visual system (Heeger and Bergen, 1995; Portilla and Simoncelli, 2000; Zhu et al., 1998). An influential example of this approach is the model by Portilla and Simoncelli (Portilla and Simoncelli, 2000), which is based on the steerable pyramid (Simoncelli and Freeman, 1995). Their model computes average filter responses as well as correlations between filter responses across space, spatial scales and orientations. New texture samples are synthesized by performing a pre-image search: an image is initialised with white noise and optimized iteratively by gradient descent such that it matches the summary statistics of the target texture. The Portilla & Simoncelli texture model has inspired numerous studies in vision research (e.g. Balas et al., 2009; Freeman et al., 2013; Freeman and Simoncelli, 2011; Okazawa et al., 2015; Rosenholtz et al., 2012) because of its good synthesis performance on many natural textures and its close relationship to the computational building blocks of the early visual system.

The success of deep learning

The modelling of more complex computational building blocks of the visual system has benefitted greatly from a revolution in the field of computer vision and machine learning. In 2012, Alex Krizhevsky and colleagues (Krizhevsky et al., 2012) outperformed the state-of-the art in object recognition in the *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) (Russakovsky et al., 2014) by a large margin using a deep convolutional neural network (CNN).

Interestingly, the representations of CNNs trained to solve large-scale object recognition appear to be generally useful for visual information processing. They transfer to other datasets and tasks (Donahue et al., 2014; Oquab et al., 2014). It has become common practice in the computer vision community to use the activations of CNNs pre-trained on object recognition as the ad-hoc feature representation to solve visual information processing tasks.

DiCarlo and co-workers showed that performance-optimized CNNs provide not only general and powerful image features for machine vision, but they also excel in predicting neural activity in primate areas V4 and IT (Cadieu et al., 2014; Yamins et al., 2014), two areas of the ventral visual stream – the part of the primate brain responsible for object recognition. Studies in humans confirmed the usefulness of CNN representations as predictors for biological vision processes. They set the state-of-the-art in predicting fMRI responses in the human ventral stream (Güçlü and Gerven, 2015; Khaligh-Razavi and Kriegeskorte, 2014) and in predicting where in images humans look (Kümmerer et al., 2015).

Thus, models from computer vision and machine learning that share only a very coarse computational architecture with the visual cortex are the best models for predicting brain responses and human visual behaviour.

A texture model based on deep neural network features

The predictive power of modern convolutional neural networks and the fruitfulness of parametric texture modelling for visual neuroscience (Freeman et al., 2013; Freeman and Simoncelli, 2011) motivated us to work on a texture model (Gatys et al., 2015) based on the feature representations of a high-performing CNN trained on object recognition (Simonyan and Zisserman, 2014). This texture model then led to the development of the widely popularised neural style transfer algorithm (Gatys et al., 2016a). Here we briefly introduce this work and discuss its relevance for vision research and the understanding of CNN representations.

The central part of this CNN-based texture algorithm is a pre-image search. Pre-image search techniques have been used before for understanding and visualising the feature representations of CNNs (Mahendran and Vedaldi, 2014; Nguyen et al., 2015; Simonyan et al., 2013; Szegedy et al., 2013) and for synthesising textures (Portilla and Simoncelli, 2000). Building on these ideas, we constrained a pre-image search by the spatial correlations between feature maps of different convolutional layers in the network (Gatys et al., 2015). The resulting textures look very realistic and many of them are indistinguishable from the originals under realistic viewing conditions (Wallis et al., 2017).

Furthermore, the texture parameters are factorized into layers of neurons that make high-level image information increasingly explicit. This property allows us to generate images with an increasing degree of naturalness (Gatys et al., 2015), which could also serve as useful stimuli for studying visual perception.

From texture modelling to style transfer

The neural network texture model forms the basis of Neural Style Transfer (Gatys et al., 2016a), an algorithm that repaints a photograph in the style of an arbitrary painting (Fig. 2). The algorithm maps the texture of the painting onto the photograph while preserving high-level features (the content) of the photograph. It can be thought of as synthesising a texture in the style of the painting with the additional constraint of matching the activations of higher-layer deep network features. Since units in higher layers are quite invariant under low-level variation, these constraints allow for a lot of flexibility in terms of drawing style.

A useful analogy for the two types of constraints – content and style – is the phase spectrum and the power spectrum of an image. It has long been known that the phase spectrum is important for recognizing the content of an image (Oppenheim and Lim, 1981), while the power spectrum captures spatial correlations in a shift-invariant manner more related to texture or style. Phase and power spectrum contain fully complementary information and can thus be arbitrarily recombined without compromises. The content and style representations obtained with convolutional neural networks are perceptually quite complementary, but not completely independent. Nevertheless, optimization methods allow us to find very good simultaneous matches to both types of constraints for many combinations of images.

The intriguing results of Neural Style Transfer raise some important questions about human perception of artistic style. Although there is no doubt that artistic style is much more complex than what the algorithm currently captures, we get quite far by modelling it purely with the texture of a painting. We generate images that 'look like' one specific painting, but show the semantic content of another image. This finding suggests that our ability to perceive and create visual arts very much relies on the interplay between texture perception and semantic image analysis like object recognition.

Moreover, we can gain new insights into the functioning of high-performing deep neural networks. For example, much of their object recognition performance stems from analysing images without maintaining the explicit spatial arrangement of object parts. The VGG network (Simonyan and Zisserman, 2014) often outputs the same class label for regular images and completely scrambled (texturised) versions of the same image (Fig. 3A). Along similar lines, we transferred the texture of an object class with a very distinct texture (e.g. Leopard) onto an image of a different class (e.g. Ford T Model) (Fig. 3B). The VGG network classified the resulting image as a leopard, although most humans would probably agree that the image shows a car.

Interestingly, texture synthesis does not appear to require learned representations, but also works with a set of random filter responses (Ustyuzhaninov et al., 2016). However, compelling style transfer requires a learned representation, presumably because it needs the invariance properties to satisfy both content and style constraints simultaneously (He et al., 2016).

Recent advances in texture synthesis and style transfer

Since the initial papers on texture synthesis (Gatys et al., 2015) and Neural Style Transfer (Gatys et al., 2016a), many researches have explored improvements and variations around the basic algorithm, in order to improve its quality, make it faster, more flexible and to apply it to more specific domains.

A first set of improvements concerns the quality of the texture output. The original texture model (Gatys et al., 2015) was not designed to capture very long-range dependencies, such as periodic patterns or large texture elements. This limitation, which is caused by the finite receptive field size of the units in the neural network, can be overcome by including additional constraints, such as matching the power spectrum (Liu et al., 2016) or shifted feature correlations (Berger and Memisevic, 2016). In addition, the texture model has been extended to model reflection (Aittala et al., 2016) and to infer photorealistic facial textures from single images (Saito et al., 2016).

Several improvements were made to Neural Style Transfer. The original algorithm does not preserve photorealism when the style is a photograph. This problem was mostly alleviated using a patch-based approach (Li and Wand, 2016a) similar to traditional patch-based texture synthesis methods (Efros et al., 1999; Wei and Levoy, 2000) with patches defined on the feature representation of the neural network instead of the raw images. The original allows the user only to specify two images and the trade-off of content versus style weight, but offers little control over the stylisation procedure. Such controls have been introduced, including spatial control (Champandard, 2016; Gatys et al., 2016b), control over the stylisation on different spatial scales and controlling the colour of the stylisation outcome (Gatys et al., 2016b). Finally, style transfer was modified to deal with the specific case of portrait transfer (Selim et al., 2016).

An obvious, but non-trivial extension of style transfer is its application to video. Two groups combined the original approach with optical flow to smoothly stylise video sequences (Anderson et al., 2016; Ruder et al., 2016). Similarly, texture synthesis can be extended to the video domain (Funke et al., 2017).

Neural Style Transfer is slow because it is based on iterative optimisation. Even on the most recent Graphics Processing Units (GPU) it takes on the order of one minute for a 512×512 image. To speed up the procedure, multiple groups have trained feed-forward networks to solve the optimisation problem (Johnson et al., 2016; Li and Wand, 2016b; Ulyanov et al., 2016). Instead of running the optimisation procedure for each image, these networks are trained to take in an image and output a stylised version of it. These networks generate images much faster (a few milliseconds per image), but fall behind in terms of quality and flexibility (only one style per trained network). However, recent works improved the general quality (Ulyanov et al., 2017; Wang et al., 2016), added spatial control (Gatys et al., 2016b) and introduced multi-style (Dumoulin et al., 2016) and even arbitrary-style networks (Chen and Schmidt, 2016; Ghiasi et al., 2017; Huang and Belongie, 2017).

Neural network feature spaces for perceptual loss functions

The key insight of Neural Style Transfer was that feature spaces of neural networks trained on object recognition allow the separation and recombination of perceptually important image features (here ‘content’ and ‘style) in an unprecedented manner. This insight touches on a fundamental problem in computer vision and image processing: to find image representations that enable the analysis, synthesis and manipulation of images with respect to perceptual variables. Closely related is the search for measures of image quality and image distortion that have a better correspondence to human perception than classical Peak Signal to Noise Ratio (PSNR) or even the improved Structural Similarity Index (SSIM) (Wang et al., 2004).

This class of problems can be summarized as generating images subject to certain perceptual constraints. While our work addressed two instances of this problem class, texture synthesis and image style transfer, there are numerous other instances with different perceptual constraints specified. They range from classical image restoration tasks, such as image super-resolution and in-painting, where missing image information is predicted to ones with more high-level constraints, for example attribute-based image synthesis, where the goal could be to render images of a specific object.

The introduction of better perceptual loss functions based on pre-trained neural network features has inspired a large body of new work in all of these areas. To achieve state of the art performance, perceptual loss functions now constitute a vital ingredient – together with adversarial losses[1] (Goodfellow et al., 2014), another recent and very popular approach to image synthesis.

Single image super-resolution has seen a considerable boost in performance by introducing the use of feature spaces of pre-trained neural networks as a measure of perceptual image quality (Bruna et al., 2015; Johnson et al., 2016; Ledig et al., 2016; Sajjadi et al., 2016). In particular, Sajjadi et al. explicitly use a texture loss on pre-trained neural network features in order to achieve state-of-the art results.

In image in-painting, Yang et al recently combined a CNN that predicts the structure of a missing part of an image with the neural patches approach (Li and Wand, 2016a) to achieve excellent perceptual results (Yang et al., 2016).

In image attribute manipulation, interesting results were achieved by mapping images into the feature spaces of pre-trained CNNs to manipulate and transfer semantic facial attributes such as the age or 'with/without glasses' (Li et al., 2016; Upchurch et al., 2016). Korshunova et al. used the VGG features to train a network that exchange face identity in portraits (Korshunova et al., 2016). Taigman et al. combine high-level features from a face recognition network with adversarial training to generate emojis from portraits while preserving facial identity (Taigman et al., 2016).

The task of attribute-based image synthesis has the longest history in image synthesis with CNNs. The idea is the following: to generate an image of a certain object, say a washing machine, one performs a pre-image search to find an image that maximises the activation of the washing machine unit in the classification layer of a CNN trained on object recognition. In their pioneering work, Simonyan and colleagues showed that this procedure indeed generates images that somewhat resemble the features of a washing machine (Simonyan et al., 2013). However, the pre-image search is under-constrained and leads to noisy, texture-like rather than natural images (Fig. 4A). These problems were addressed by regularising the pre-image search with natural-image priors, such as constraints on the total variation of the images (Mahendran and Vedaldi, 2014; Yosinski et al., 2015). More recently, Nguyen et al. (Nguyen et al., 2016a, 2016b) managed to generate diverse and fully realistic natural images of certain object classes (Fig 4B) by using

[1] The idea behind adversarial losses is to train a second, so-called adversarial network, whose task is to tell apart synthesised images from originals. By training the generator and the adversarial network concurrently, one can ensure that the generator produces realistically looking samples

a generator network that is trained to invert CNN representations (Dosovitskiy and Brox, 2016) as a regulariser for the pre-image search.

The prominent and widely popularised Deep Dream algorithm from Google also built on this principle of activation maximization (Mordvintsev et al., 2015). Deep Dream synthesises or manipulates images to maximize the activation of specific feature maps. Since these features encode semantic properties, such as objects or parts of objects or scenes, the synthesised images look like hallucinations of these semantic entities (Fig 4C). The huge impact these images generated in the digital community underscore the vast potential of CNN features as a tool for digital artists to create perceptually exciting image experiences.

Finally, recent work on sketch inversion gives further insights in how pre-trained CNN features can serve assisting the creative process in digital image creation (Güçlütürk et al., 2016; Sangkloy et al., 2016).

In summary, feature spaces that correspond to human perception have enabled a wide range of improvements in image editing and synthesis, which range from low-level tasks such as image super resolution to high-level tasks such as attribute-based image synthesis.

Conclusions

We have seen how ideas originating in visual neuroscience have had a substantial impact on the computer vision community. But can neuroscience learn something from these developments? The major advances in image synthesis and editing discussed in this review became possible because current CNNs like the VGG network are good models of human perception beyond the specific task that they have been originally trained for (Donahue et al., 2014; Kümmerer et al., 2015; Oquab et al., 2014).

In our opinion, the power of current CNNs to model natural perception renders them the most useful models for understanding the visual system in the brain. Importantly, this claim does not build on the superficial similarity between the CNNs and the neural anatomy in the visual pathway. At the connectomics level, current CNNs are not more similar to the visual pathway than the Neocognitron, a multi-layer neural network originally proposed by Fukushima (Fukushima, 1980), or later models like H-MAX (Riesenhuber and Poggio, 1999).

Naive assessment of “neural similarity” can be highly misleading for understanding neural function. Despite their high-level similarity with modern CNNs, neither Neocognitron nor H-MAX performs well on real-world tasks. They also fail at predicting neural responses (Yamins et al., 2014). One meaningful way to define “neural similarity” between artificial and biological neural networks is system identification (Cadieu et al., 2014; Güçlü and Gerven, 2015; Yamins et al., 2014). Establishing such a correspondence between CNNs and biological neural networks is a promising start, but understanding human perception will take much more. An obvious further direction is to combine system identification with perceptually meaningful image synthesis methods. Such an approach could generate stimuli to explore the response properties of biological neurons and their impact on perception in a directed manner.

We believe that the insights we gain from advances of CNNs have profound implications for Computational Neuroscience. At the level at which we currently define understanding in Neuroscience, the Neocognitron and the VGG network are largely equivalent. However, the vast difference in performance between the two models demonstrates that our naïve judgements of “neural similarity” can be of little relevance for assessing the effectiveness of neural computation. Even for such simple neural networks, for which the entire connectome is known and simultaneous recordings of all the neurons can be done efficiently, we do not have any theory yet how to explain the striking difference in their behaviour.

If we want to understand the brain, the explanation of behavioural differences between artificial neural networks is an extremely important exercise. It allows us to get a better idea of what constitutes an explanation of neural network function and what we could hope to accomplish if there were no experimental limitations. Explaining the difference between a VGG network and the Neocognitron is just one example of many questions that a theory of neural network computation should be able to answer. High-performing artificial neural networks offer a huge opportunity for Computational Neuroscience, because they are great for exploring how effectively our theoretical tools can facilitate an understanding of the neural basis of behaviour.

Funding

(a) German Research Foundation (CRC 1233 on Robust Vision)
(b) German Research Foundation (DFG grant number EC 479/1-1)
(c) IARPA: "the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government."

Bibliography

Adelson, E.H., Bergen, J.R., 1991. The plenoptic function and the elements of early vision.

Aittala, M., Aila, T., Lehtinen, J., 2016. Reflectance Modeling by Neural Texture Synthesis. ACM Trans. Graph. 35.

Anderson, A.G., Berg, C.P., Mossing, D.P., Olshausen, B.A., 2016. DeepMovie: Using Optical Flow and Deep Neural Networks to Stylize Movies. ArXiv Prepr. ArXiv160508153.

Balas, B., Nakano, L., Rosenholtz, R., 2009. A summary-statistic representation in peripheral vision explains visual crowding. J. Vis. 9, 13.

Berger, G., Memisevic, R., 2016. Incorporating long-range consistency in CNN-based texture generation. ArXiv Prepr. ArXiv160601286.

Bruna, J., Sprechmann, P., LeCun, Y., 2015. Super-Resolution with Deep Convolutional Sufficient Statistics. ArXiv151105666 Cs.

**Cadieu, C.F., Hong, H., Yamins, D.L.K., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., DiCarlo, J.J., 2014. Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. PLoS Comput Biol 10, e1003963. doi:10.1371/journal.pcbi.1003963

Showed that representations learned by deep networks are good feature spaces for explaining neural responses in the visual ventral stream of the brain.

Champandard, A.J., 2016. Semantic Style Transfer and Turning Two-Bit Doodles into Fine Artworks. ArXiv160301768 Cs.

Chen, T.Q., Schmidt, M., 2016. Fast Patch-based Style Transfer of Arbitrary Style. ArXiv161204337 Cs.

**Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T., 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition., in: Icml. pp. 647–655.

Showed that representations learned by deep convolutional neural networks transfer to other tasks and datasets the network was not originally trained on.

Dosovitskiy, A., Brox, T., 2016. Generating images with perceptual similarity metrics based on deep networks, in: Advances in Neural Information Processing Systems. pp. 658–666.

Dubuc, B., Zucker, S.W., 2001. Complexity, confusion, and perceptual grouping. Part I: The curve-like representation. Int. J. Comput. Vis. 42, 55–82.

Dumoulin, V., Shlens, J., Kudlur, M., 2016. A Learned Representation For Artistic Style. ArXiv161007629 Cs.

Efros, A., Leung, T.K., others, 1999. Texture synthesis by non-parametric sampling, in: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference On. IEEE, pp. 1033–1038.

Freeman, J., Simoncelli, E.P., 2011. Metamers of the ventral stream. Nat. Neurosci. 14, 1195–1201. doi:10.1038/nn.2889

Freeman, J., Ziemba, C.M., Heeger, D.J., Simoncelli, E.P., Movshon, J.A., 2013. A functional and perceptual signature of the second visual area in primates. Nat. Neurosci. 16, 974–981. doi:10.1038/nn.3402

Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. 36, 193–202. doi:10.1007/BF00344251

Funke, C.M., Gatys, L.A., Ecker, A.S., Bethge, M., 2017. Synthesising Dynamic Textures using Convolutional Neural Networks. ArXiv170207006 Cs.

* Gatys, L., Ecker, A.S., Bethge, M., 2015. Texture Synthesis Using Convolutional Neural Networks, in: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp. 262–270.

A parametric texture model that produces realistically looking textures using summary statistics of deep network activations.

** Gatys, L.A., Ecker, A.S., Bethge, M., 2016a. Image style transfer using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2414–2423.

Showed that deep network feature spaces can serve as powerful loss functions for image synthesis and introduced a content and a style representation, which are largely independent.

Gatys, L.A., Ecker, A.S., Bethge, M., Hertzmann, A., Shechtman, E., 2016b. Controlling Perceptual Factors in Neural Style Transfer. ArXiv Prepr. ArXiv161107865.

Ghiasi, G., Lee, H., Kudlur, M., Dumoulin, V., Shlens, J., 2017. Exploring the structure of a real-time, arbitrary neural artistic stylization network. ArXiv170506830 Cs.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 2672–2680.

Güçlü, U., Gerven, M.A.J. van, 2015. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. J. Neurosci. 35, 10005–10014. doi:10.1523/JNEUROSCI.5023-14.2015

Güçlütürk, Y., Güçlü, U., van Lier, R., van Gerven, M.A.J., 2016. Convolutional Sketch Inversion. ArXiv160603073 Cs 9913, 810–824. doi:10.1007/978-3-319-46604-0_56

He, K., Wang, Y., Hopcroft, J., 2016. A Powerful Generative Model Using Random Weights for the Deep Image Representation, in: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 29. Curran Associates, Inc., pp. 631–639.

Heeger, D.J., Bergen, J.R., 1995. Pyramid-based Texture Analysis/Synthesis, in: Proceedings of the 22Nd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '95. ACM, New York, NY, USA, pp. 229–238. doi:10.1145/218380.218446

Huang, X., Belongie, S., 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. ArXiv170306868 Cs.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Lecture Notes in Computer Science. Springer International Publishing, pp. 694–711. doi:10.1007/978-3-319-46475-6_43

Julesz, B., 1984. A brief outline of the texton theory of human vision. Trends Neurosci. 7, 41–45.

Julesz, B., 1962. Visual Pattern Discrimination. IRE Trans. Inf. Theory 8, 84–92. doi:10.1109/TIT.1962.1057698

Julesz, B., Gilbert, E.N., Shepp, L.A., Frisch, H.L., 1973. Inability of Humans to Discriminate between Visual Textures That Agree in Second-Order Statistics—Revisited. Perception 2, 391–405. doi:10.1068/p020391

Julesz, B., Gilbert, E.N., Victor, J.D., 1978. Visual discrimination of textures with identical third-order statistics. Biol. Cybern. 31, 137–140.

Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput Biol 10, e1003915. doi:10.1371/journal.pcbi.1003915

Korshunova, I., Shi, W., Dambre, J., Theis, L., 2016. Fast Face-swap Using Convolutional Neural Networks. ArXiv161109577 Cs.

**Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems. pp. 1097–1105.

The work that started the deep learning revolution by cutting the error rate in the ImageNet Large Scale Visual Recognition Challenge in half by using a convolutional neural network.

Kümmerer, M., Theis, L., Bethge, M., 2015. Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet, in: ICLR Workshop.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W., 2016. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. ArXiv160904802 Cs Stat.

Li, C., Wand, M., 2016a. Combining markov random fields and convolutional neural networks for image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2479–2486.

Li, C., Wand, M., 2016b. Precomputed real-time texture synthesis with markovian generative adversarial networks, in: European Conference on Computer Vision. Springer, pp. 702–716.

Li, M., Zuo, W., Zhang, D., 2016. Convolutional Network for Attribute-driven and Identity-preserving Human Face Generation. ArXiv160806434 Cs.

Liu, G., Gousseau, Y., Xia, G.-S., 2016. Texture Synthesis Through Convolutional Neural Networks and Spectrum Constraints. ArXiv160501141 Cs.

Mahendran, A., Vedaldi, A., 2014. Understanding Deep Image Representations by Inverting Them. ArXiv14120035 Cs.

*Mordvintsev, Alexander, Olah, Christopher, Tyka, Mike, 2015. Research Blog: Inceptionism: Going Deeper into Neural Networks [WWW Document]. URL https://web.archive.org/web/20150703064823/http://googleresearch.blogspot.co.uk/2015/06/inceptionism-going-deeper-into-neural.html (accessed 2.16.17).

Widely popularized algorithm that first demonstrated the potential applicability of CNNs for digital art.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J., 2016a. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, in: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 29. Curran Associates, Inc., pp. 3387–3395.

Nguyen, A., Yosinski, J., Bengio, Y., Dosovitskiy, A., Clune, J., 2016b. Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space. ArXiv161200005 Cs.

Nguyen, A., Yosinski, J., Clune, J., 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 427–436.

Okazawa, G., Tajima, S., Komatsu, H., 2015. Image statistics underlying natural texture selectivity of neurons in macaque V4. Proc. Natl. Acad. Sci. 112, E351–E360. doi:10.1073/pnas.1415146112

Oppenheim, A.V., Lim, J.S., 1981. The importance of phase in signals. Proc. IEEE 69, 529–541. doi:10.1109/PROC.1981.12022

Oquab, M., Bottou, L., Laptev, I., Sivic, J., 2014. Learning and transferring mid-level image representations using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1717–1724.

Portilla, J., Simoncelli, E.P., 2000. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. Int. J. Comput. Vis. 40, 49–70. doi:10.1023/A:1026553619983

Purpura, K.P., Victor, J.D., Katz, E., 1994. Striate cortex extracts higher-order spatial correlations from visual textures. Proc. Natl. Acad. Sci. U. S. A. 91, 8482–8486.

Riesenhuber, M., Poggio, T., 1999. Hierarchical models of object recognition in cortex. Nat. Neurosci. 2, 1019–1025. doi:10.1038/14819

Rosenholtz, R., Huang, J., Raj, A., Balas, B.J., Ilie, L., 2012. A summary statistic representation in peripheral vision explains visual search. J. Vis. 12, 14.

Ruder, M., Dosovitskiy, A., Brox, T., 2016. Artistic style transfer for videos, in: German Conference on Pattern Recognition. Springer, pp. 26–36.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2014. ImageNet Large Scale Visual Recognition Challenge. ArXiv14090575 Cs.

Saito, S., Wei, L., Hu, L., Nagano, K., Li, H., 2016. Photorealistic Facial Texture Inference Using Deep Neural Networks. ArXiv161200523 Cs.

Sajjadi, M.S.M., Schölkopf, B., Hirsch, M., 2016. EnhanceNet: Single Image Super-Resolution through Automated Texture Synthesis. ArXiv161207919 Cs.

Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J., 2016. Scribbler: Controlling Deep Image Synthesis with Sketch and Color. ArXiv161200835 Cs.

Selim, A., Elgharib, M., Doyle, L., 2016. Painting style transfer for head portraits using convolutional neural networks. ACM Trans. Graph. TOG 35, 129.

Simoncelli, E.P., Freeman, W.T., 1995. The steerable pyramid: A flexible architecture for multi-scale derivative computation, in: Image Processing, International Conference On. IEEE Computer Society, pp. 3444–3444.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. ArXiv13126034 Cs.

*Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv14091556 Cs.

A very popular and publically available convolutional neural network with a particularly simple architecture trained on large-scale object recognition.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2013. Intriguing properties of neural networks. ArXiv13126199 Cs.

Taigman, Y., Polyak, A., Wolf, L., 2016. Unsupervised Cross-Domain Image Generation. ArXiv161102200 Cs.

Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V., 2016. Texture Networks: Feed-forward Synthesis of Textures and Stylized Images. ArXiv160303417 Cs.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2017. Improved Texture Networks: Maximizing Quality and Diversity in Feed-forward Stylization and Texture Synthesis. ArXiv Prepr. ArXiv170102096.

Upchurch, P., Gardner, J., Bala, K., Pless, R., Snavely, N., Weinberger, K., 2016. Deep Feature Interpolation for Image Content Changes. ArXiv161105507 Cs.

Ustyuzhaninov, I., Brendel, W., Gatys, L.A., Bethge, M., 2016. Texture synthesis using shallow convolutional networks with random filters. ArXiv Prepr. ArXiv160600021.

Wallis, T.S.A., Funke, C.M., Ecker, A.S., Gatys, L.A., Wichmann, F.A., Bethge, M., 2017. A parametric texture model based on deep convolutional features closely matches texture appearance for humans. bioRxiv 165761. doi:10.1101/165761

Wang, X., Oxholm, G., Zhang, D., Wang, Y.-F., 2016. Multimodal Transfer: A Hierarchical Deep Convolutional Neural Network for Fast Artistic Style Transfer. ArXiv Prepr. ArXiv161201895.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13, 600–612.

Wei, L.-Y., Levoy, M., 2000. Fast texture synthesis using tree-structured vector quantization, in: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. ACM Press/Addison-Wesley Publishing Co., pp. 479–488.

Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. 201403112. doi:10.1073/pnas.1403112111

Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H., 2016. High-Resolution Image Inpainting using Multi-Scale Neural Patch Synthesis. ArXiv161109969 Cs.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H., 2015. Understanding Neural Networks Through Deep Visualization. ArXiv150606579 Cs.

Yu, Y., Schmid, A.M., Victor, J.D., 2015. Visual processing of informative multipoint correlations arises primarily in V2. eLife 4, e06604.

Zhu, S.C., Wu, Y., Mumford, D., 1998. Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. Int. J. Comput. Vis. 27, 107–126.

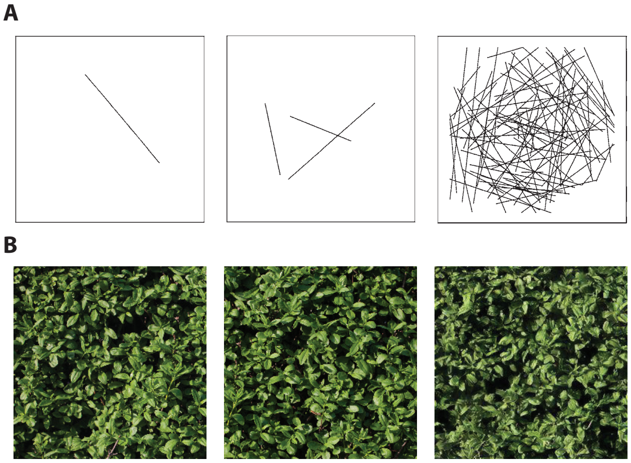

Figure 1: **A,** Demonstration of texture perception. A few elements are perceived as individual elements (left, middle). When enough similar elements are close together, we perceive them as one texture entity ('stuff', right) rather than a collection of individual things. (from Dubuc and Zucker, 2001). **B,** Several samples from the same texture. The rightmost sample was generated using our CNN texture model (Gatys et al., 2015) .

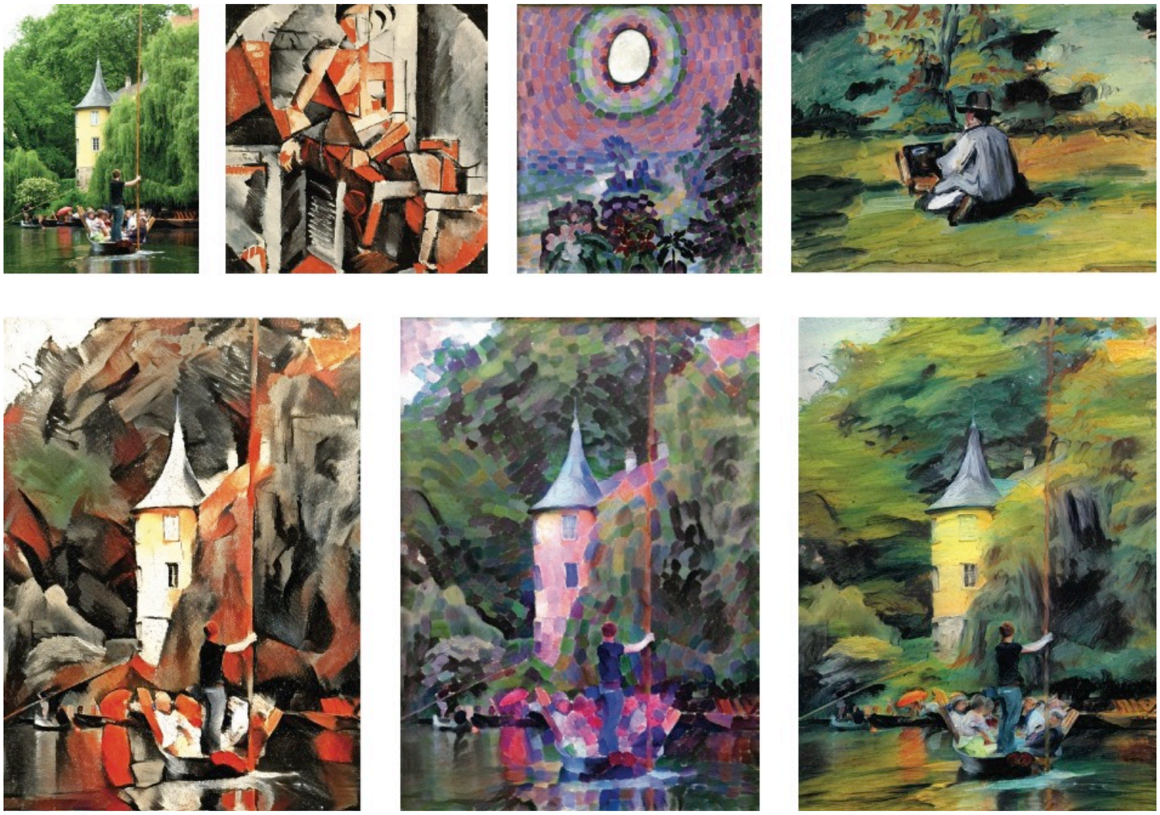

Figure 2: Demonstration of the Neural Style Transfer algorithm (Gatys et al., 2016a). A photograph (top left) is transformed to reproduce the style of several different paintings.

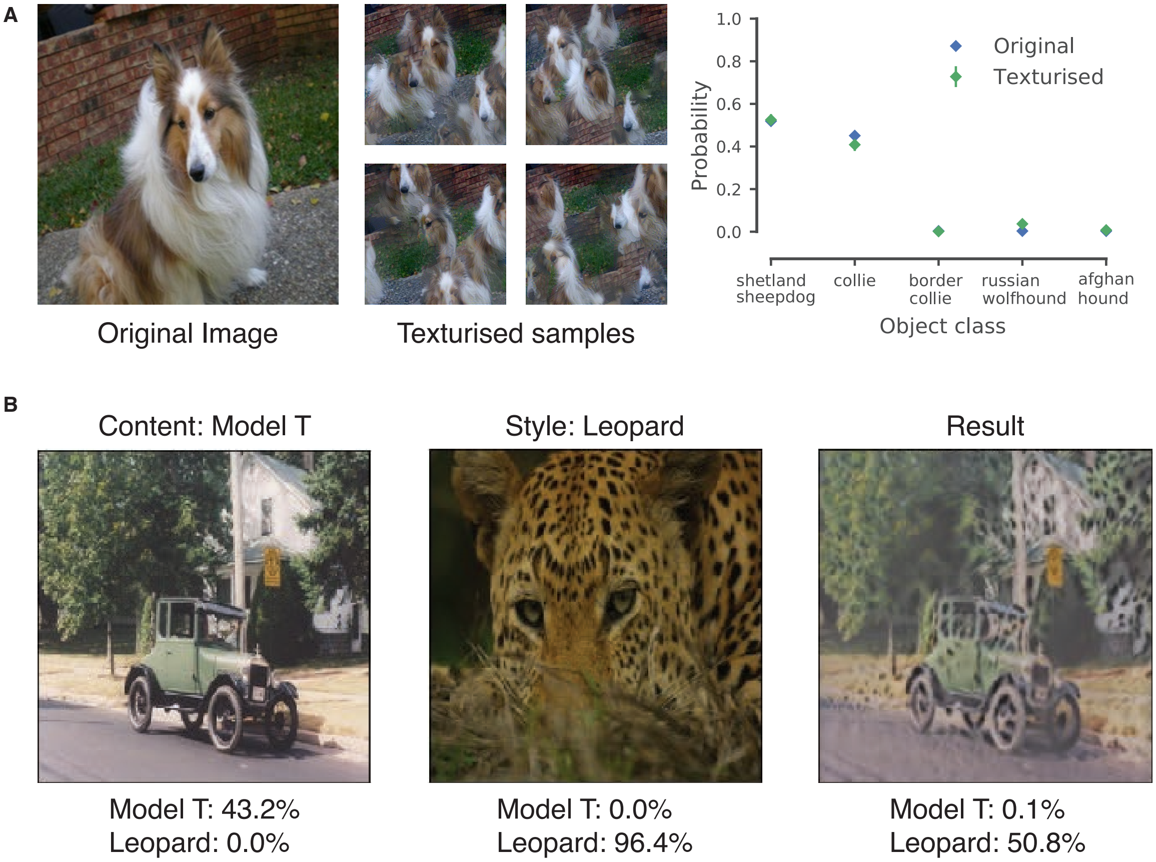

Figure 3: **A,** Texturised samples of a non-texture image are classified the same as the source image. We generated 19 texturised samples (middle) of an image displaying a dog (left) using the CNN texture model (Gatys et al., 2015). We classified both the original and the texturised samples using the VGG network (Simonyan and Zisserman, 2014). We plot the probability for the top-5 object classes predicted in response to the original image (left). Additionally, we plot and the mean and the SEM of the classification of the texture samples (left). Even though the spatial integrity of the dog is completely lost in the texture samples, they are classified extremely similar to the original image (left). **B,** We perform colour-preserving style transfer (Gatys et al., 2016b) from an image showing a leopard (middle) onto an image showing a car (left). While the VGG network (Simonyan and Zisserman, 2014) classifies the content image correctly as a "Ford Model T", it classifies the image after the style transfer (right) as a leopard.

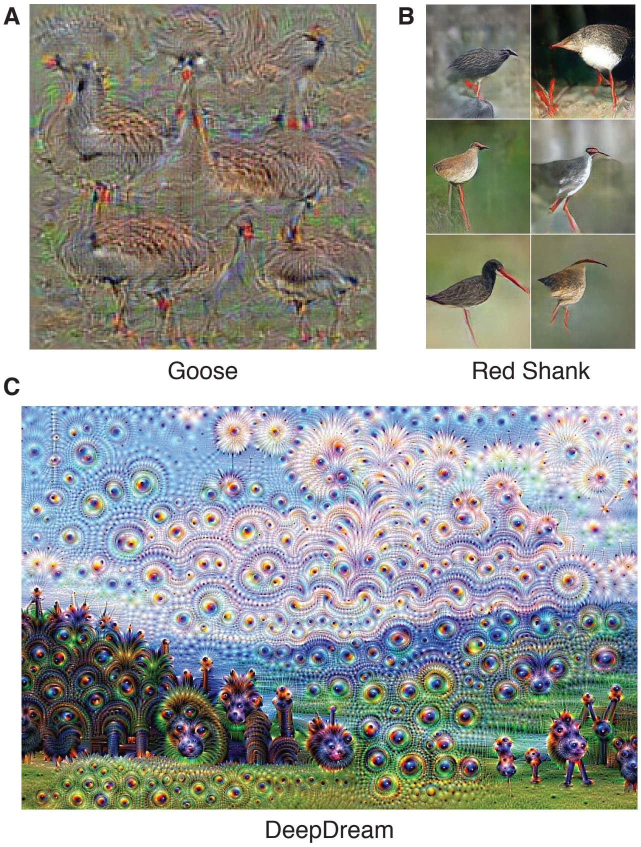

Figure 4: **A,** Image that maximises the activation of the 'Goose' unit in a CNN found with naive gradient descent without regularisation (from Simonyan et al., 2013). **B,** Images that maximise the activation of the 'Red Shank' unit in a CNN found using regularisation via a generator network (from Nguyen et al., 2016b). **C,** Example of a photograph that was artistically modified using the Deep Dream algorithm (Mordvintsev et al., 2015).