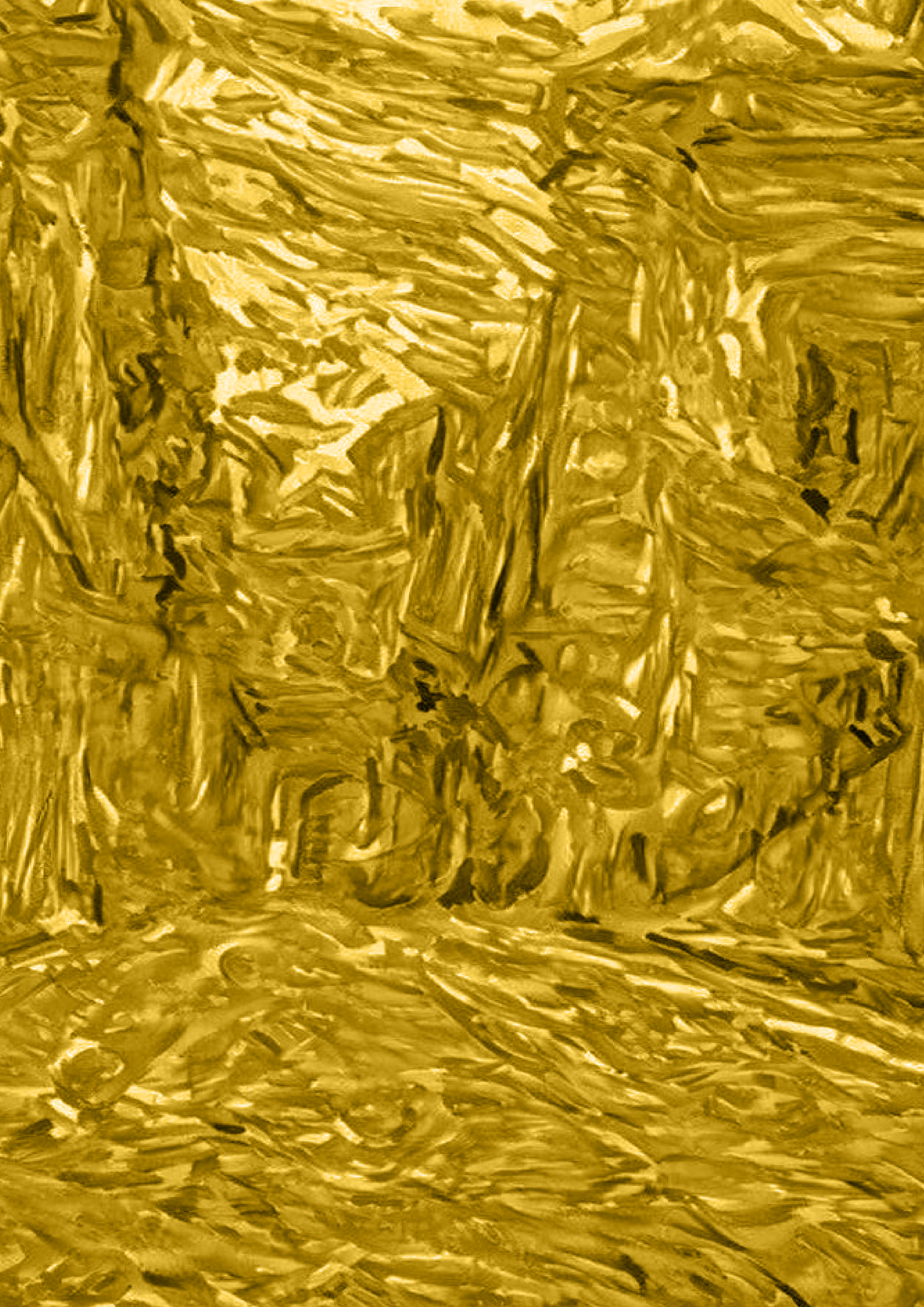

IMAGE SYNTHESIS AS A METHOD OF KNOWLEDGE PRODUCTION IN ART HISTORY

MATTHIAS WRIGHT, BJÖRN OMMER

ABSTRACT | Digital images enable us to virtually assemble, group, and rearrange works of art as image datasets. The highly complex similarities and dissimilarities between data points in an image dataset can be analyzed. Understanding the meaning of computationally defined similarities and dissimilarities, however, requires disentangling the representations learned by the computer in the process. By utilizing generative methods from deep learning, we aim to design a new methodology for the analysis and interpretation of digital images. Building on refined methods of disentanglement from computer science, our goal is to establish the synthetic image as a novel means of knowledge production in art history.

KEYWORDS | machine learning, computer vision, deep learning, image synthesis, artistic style

Computer Vision

The field of computer vision has its origin in the early 1970s.[1] In the beginning, it was merely intended to be the visual perception component of a system that mimics human intelligence.[2] Some of the early pioneers of artificial intelligence believed that creating this component would be fairly easy compared to problems such as higher-level reasoning or planning.[3] In 1966, Marvin Minsky even asked an undergraduate student to "spend the summer linking a camera to a computer and getting the computer to describe what it saw".[4]

From its beginnings in the early 1970s up until the 1990s computer vision research was mostly concerned with perception – describing objects or scenes in images.[5] However, during the 1990s, computer vision and computer graphics became more and more intertwined,[6] a trend that continued to the present day.

Over the past decade, the field of computer vision has become increasingly dominated by deep learning, a class of machine learning methods that are "representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level".[7]

Representations

The concept of representations is central for the field of computer vision. In the words of David Marr: "A representation is a formal system for making explicit certain entities or types of information, together with a specification of how the system does this".[8] The result of using a representation to describe some entity is then called a description of the entity in that representation.[9] We use representations every day, sometimes even without knowing it. For example, the same number may be represented in different numeral systems.[10]

The notion of representations is powerful because how information is represented "can greatly affect how easy it is to do different things with it".[11] This is not just true for computer vision but also for mathematics. Just consider the eigendecomposition of a matrix. If certain conditions are satisfied, a square matrix can be represented as the product of three matrices. This representation exhibits information about the functional properties of the matrix not apparent from the canonical matrix representation.[12]

In the context of computer vision, we usually deal with data arising from the complicated interaction of many factors. For example, an image consists of the interaction between one or several light sources, the shapes of the objects, the material of the surfaces that occur in the image, and the viewpoint[13]. If our task is object classification, we would want a representation of the image that is invariant to light and viewpoint but not to object shape or material. This is because a dog is always a dog, no matter how bright the image is. The viewpoint from which the dog is depicted, should not affect the classification result either.

However, our choice of invariant features generally depends on the task we are trying to accomplish. If our goal was to determine whether or not an image was taken by day or by night, light would suddenly become an important factor.

Unfortunately, in many cases we do not know a priori which set of features and variations will be relevant for our task.[14] Therefore, the most robust approach is to "disentangle as many factors as possible, discarding as little information about the data as is practical".[15]

The definition of a disentangled representation is based on three criteria: modularity, compactness, and explicitness.[16] A representation is modular if each component of the representation contains information about at most one factor.[17] A representation is compact if a given factor is associated with only one or a few components of the representation.[18] A representation is explicit if there is a simple mapping from the component to the value of a factor.[19]

Computer Vision and Art History

A great advantage of digital images is their potential to bring large numbers of artifacts together virtually in order to then easily link them to related samples, to flexibly rearrange them, or simply to order them in database systems. Much like in Aby Warburg's mnemosyne atlas, digital images are therefore constantly being brought into relation to another. However, relations and similarities or dissimilarities between artworks are based on potentially fairly abstract representations. Especially when computers establish such relations.

In recent years, there has been a surge of deep learning approaches that are generative in nature.[20] These methods allow the direct visualization of the abstract representations that they learn. A relevant example of this is Neural Style Transfer, which refers to a class of image synthesis algorithms that aim to render an image into the style of a given artwork. See fig. 1 and fig. 2 for example images.

The original method was proposed by Gatys, Ecker, and Bethge[21] and consisted of an iterative optimization procedure, which optimized a combination of two objective functions. The first objective function ensures that the stylized image still contains the content from the original image, which is measured by the learned image representations of a convolutional neural network that was trained for image classification. The second objective function encourages the stylized image to have a similar style to the given artwork, which is measured using the Gramian matrices of the learned image representations.

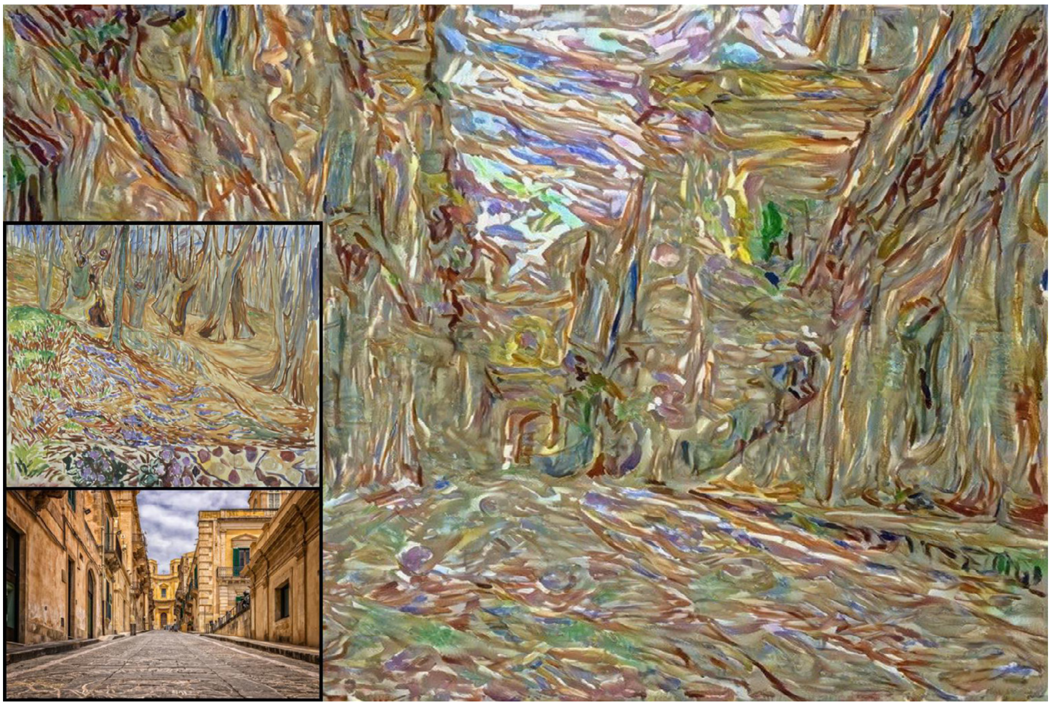

Figure 1. An image of a road rendered in the style of "Spring in the Elm Forest" by Edvard Munch; rendering by the authors; 2020.

Several methods have been proposed that employ a neural network to approximate the optimization objective from Gatys, Ecker, and Bethge[22] for a specific artwork.[23] The underlying problem that Neural Style Transfer methods aim to solve is a disentangling of style and content. The algorithm needs to extract the semantic content from the input image and render it into the style that was distilled from the artwork. This problem is highly relevant, even beyond the area of Neural Style Transfer. Imagine our goal is to group a large collection of different artworks with respect to their content. This is not a trivial task, because the same object might look very different when depicted in two different styles. Just compare a portrait painted by Picasso with a portrait from Da Vinci. Image representations for those artworks that decompose into separate style and content representations would enable us to find semantic correspondence between artworks across a wide range of different styles. Techniques from Neural Style Transfer[24] have also been employed for controlled image synthesis.[25] The proposed method learned a disentangling of high-level attributes (e.g. of human faces) as well as stochastic variation of low-level features.[26]
This project will work with neural networks that synthetically generate digital images to explain the representations they have learned for art collections. These representations can give novel insights into cultural artifacts that are not tangible through human natural language.[27] The generated synthetic digital images establish a new means of access to concepts in collections of digital or digitized art by distillation. Consequently, our goal in this project is to challenge the way art history views the digital image. The digital image should convert towards an epistemic instrument. Rather than only being the object of an art historical analysis, we will empower synthetic digital images to become a valuable tool for the analysis process. The project tackles the hermeneutic questions of reading not only a 'computer generated image' but the underlying manifold.

NOTES

[1] Richard Szeliski, *Computer Vision: Algorithms and Applications* (Berlin: Springer, 2010), 11.

[2] Szeliski, *Computer Vision: Algorithms and Applications,* 11.

[3] Szeliski, *Computer Vision: Algorithms and Applications,* 11.

[4] Margaret Boden, *Mind as Machine: A History of Cognitive Science* (Oxford: Oxford University Press, 2006), 781.

[5] Richard Szeliski, *Computer Vision: Algorithms and Applications* (Berlin: Springer, 2010), 10–18.

[6] Szeliski, *Computer Vision: Algorithms and Applications*, 17.

[7] Yann LeCun, Yoshua Bengio and Geoffrey Hinton, "Deep learning," *Nature* 521 (2015): 436. https://www.nature.com/articles/nature14539

[8] David Marr, *Vision*: *A Computational Investigation into the Human Representation and Processing of Visual Information*. (Cambridge: MIT Press, 2010), 20.

[9] Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, 20.

[10] Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual*, 20.

[11] Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual*, 21.

[12] Ian J. Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning* (Cambridge: MIT Press, 2016), 42.

[13] Yoshua Bengio, "Deep Learning of Representations: Looking Forward," in *Statistical Language and Speech Processing*, ed. Carlos Martín-Vide, Matthew Purver and Senja Pollak (Berlin, Heidelberg: Springer, 2013), 19.

[14] Bengio, "Deep Learning of Representations: Looking Forward," 19.

[15] Bengio, "Deep Learning of Representations: Looking Forward," 20.

[16] Cian Eastwood and Christopher K.I Williams, "A framework for the quantitative evaluation of disentangled representations." In *International Conference on Learning Representations (ICLR),* Vancouver, April 30–May 3, 2018. https://openreview.net/pdf?id=By-7dz-AZ. See also Karl Ridgeway, and Michael C. Mozer, "Learning deep disentangled embeddings with the f-statistic loss." In *Conference on Neural Information Processing Systems (NIPS)*, Montréal, December 2–8, 2018. a

[17] Ridgeway and Mozer, "Learning deep disentangled embedding with the f-statistic loss".

[18] Ridgeway and Mozer, "Learning deep disentangled embedding with the f-statistic loss".

[19] Ridgeway and Mozer, "Learning deep disentangled embedding with the f-statistic loss".

[20] Diederik P. Kingma, and Max Welling, "Auto-Encoding Variational Bayes." In *International Conference on Learning Representations (ICLR), Banff, April 14*–16, 2014). https://arxiv.org/abs/1312.6114. See also Ian Goodfellow, et al. "Generative Adversarial Nets." In *Conference on Neural Information Processing Systems (NIPS),* Montréal, December 8–13, 2014. https://papers.nips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[21] Leon A Gatys, Alexander S. Ecker, and Matthias Bethge, "Image Style Transfer Using Convolutional Neural Networks." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, June 26–July 1, 2016, 2414-2423. https://ieeexplore.ieee.org/document/7780634.

[22] Gatys, Ecker, and Bethge, "Image Style Transfer Using Convolutional Neural Networks".

[23] Justin Johnson, Alexandre Alahi, and Fei-Fei Li, "Perceptual losses for real-time style transfer and super-resolution." In *European Conference on Computer Vision (ECCV), Amsterdam, October 8*–16, 2016. https://arxiv.org/abs/1603.08155. See also Dmitry Ulyanov, Vadim Lebedev, Vedaldi, Andrea, and Lempitsky, Victor. "Texture networks: Feed-forward synthesis of textures and stylized images." In Proceedings of Machine Learning Research, pp. 1349–1357, New York City, 2016. http://proceedings.mlr.press/v48/ulyanov16.html.

[24] Xun Huang and Serge Belongie, "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization." In *International Conference on Computer Vision (ICCV), Venice,* October 22–29, 2017. https://openaccess.thecvf.com/content_ICCV_2017/papers/Huang_Arbitrary_Style_Transfer_ICCV_2017_paper.pdf.

[25] Tero Karras, Laine Samuli, and Timo Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks." In *Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, June 16*–20, 2019. https://arxiv.org/abs/1812.04948.

[26] Karras, Laine, and Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks".

[27] Lev Manovich, "Computer vision, human senses, and language of art." *AI & Society* (2020), paragraph "Computer vision and digital humanities". Accessed December 10, 2020. https://doi.org/10.100700146-020-01094-9.

BIBLIOGRAPHY

Bengio, Yoshua. “Deep Learning of Representations: Looking Forward.” In *Statistical Language and Speech Processing. SLSP 2013. Lecture Notes in Computer Science, vol 7978, edited by Adrian-Horia Dediu, Carlos Martín-Vide, Ruslan Mitkov, and Bianca Truthe, 1–37. Berlin, Heidelberg: Springer, 2013.*

Boden, Margaret. *Mind as Machine: A History of Cognitive Science.* Oxford: Oxford University Press, 2006.

Eastwood, Cian, and Christopher K. I. Williams. 2018. “A framework for the quantitative evaluation of disentangled representations.” In *Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, April 30–May 3, 2018.* https://openreview.net/pdf?id=By-7dz-AZ.

Gatys, Leon A., Alexander S. Ecker and Matthias Bethge. 2016. “Image Style Transfer Using Convolutional Neural Networks.” In *Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, June 26—July 1, 2016, 2414-2423.* https://ieeexplore.ieee.org/document/7780634.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. 2014. “Generative Adversarial Nets.” In *Proceedings of the Conference on Neural Information Processing Systems (NIPS), Montréal, December 8–13, 2014.* https://papers.nips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Goodfellow, Ian J., Yoshua Bengio and Aaron Courville. *Deep Learning.* Cambridge: The MIT Press, 2016.

Huang, Xun, and Serge Belongie. 2017. “Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization.” In *Proceedings of the International Conference on Computer Vision (ICCV), Venice, October 22–29, 2017.* https://openaccess.thecvf.com/content_ICCV_2017/papers/Huang_Arbitrary_Style_Transfer_ICCV_2017_paper.pdf.

Johnson, Justin, Alexandre Alahi and Fei-Fei Li. 2016. “Perceptual losses for real-time style transfer and super-resolution.” In *Proceedings of the European Conference on Computer Vision (ECCV), European Conference on Computer Vision (ECCV), Amsterdam, October 8–16, 2016. https://arxiv.org/abs/1603.08155.*

Karras, Tero, Samuli Laine and Timo Aila. 2019. “A Style-Based Generator Architecture for Generative Adversarial Networks.” In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, June 16–20, 2019. https://arxiv.org/abs/1812.04948.*

Kingma, Diederik P., and Max Welling. 2014. “Auto-Encoding Variational Bayes.” In *Proceedings of the International Conference on Learning Representations (ICLR), Banff, April 14–16, 2014. https://arxiv.org/abs/1312.6114.*

LeCun, Yann, Yoshua Bengio and Geoffrey Hinton. “Deep learning.” Nature 521 (2015): 436–444. https://www.nature.com/articles/nature14539.

Manovich, Lev. “Computer vision, human senses, and language of art.” *AI & Society* (2020). Accessed December 10, 2020. https://doi.org/10.1007/s00146-020-01094-9.

Marr, David. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge: The MIT Press, 2010.

Ridgeway, Karl, and Michael C. Mozer. 2018. “Learning deep disentangled embeddings with the f-statistic loss.” In *Conference on Neural Information Processing Systems (NIPS), Montréal, December 2–8, 2018.* https://papers.nips.cc/paper/2018/file/2b24d495052a8ce66358eb576b8912c8-Paper.pdf.

Szeliski, Richard. *Computer Vision: Algorithms and Applications.* Berlin: Springer, 2010.

Ulyanov, Dmitry, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. “Texture networks: Feed-forward synthesis of textures and stylized images.” In *Proceedings of Machine Learning Research*, 1349–1357. New York City, 2016.

MATTHIAS WRIGHT is a Ph.D. student in the computer vision group at Heidelberg University. He previously completed his M.Sc. in Computer Science from the University of Bath, United Kingdom. His work focuses on neural style transfer and image synthesis.

Correspondence e-mail: matthias.wright@iwr.uni-heidelberg.de

BJÖRN OMMER received a diploma in computer science from the University of Bonn, Germany and a Ph.D. from ETH Zurich. Afterwards he held a postdoctoral position at the University of California at Berkeley. He then joined the Department of Mathematics and Computer Science at Heidelberg University as a professor where he heads the computer vision group. His research interests are in computer vision, machine learning, and cognitive science. He is an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence.

Correspondence e-mail: bjoern.ommer@iwr.uni-heidelberg.de