https://doi.org/10.2352/EI.2022.34.13.CVAA-186

Artist-specific style transfer for semantic segmentation of paintings: The value of large corpora of surrogate artworks

Thomas Heitzinger†, Matthias Wödlinger†; Technical University of Vienna, Karlsplatz 13, Vienna Austria
David G. Stork; Consultant, Portola Valley, CA 94028 USA

Abstract

Deep neural networks for semantic segmentation have recently outperformed other methods for natural images, partly due to the abundance of training data for this case. However, applying these networks to pictures from a different domain often leads to a significant drop in accuracy. Fine art paintings for highly stylized works, such as from Cubism or Expressionism, in particular, are challenging due to large deviations in shape and texture of certain objects when compared to natural images. In this paper, we demonstrate that style transfer can be used as a form of data augmentation during the training of CNN based semantic segmentation models to improve the accuracy of semantic segmentation models in art pieces of a specific artist. For this, we pick a selection of paintings from a specific style for the painters Egon Schiele, Vincent Van Gogh, Pablo Picasso and Willem de Kooning, create stylized training dataset by transferring artist-specific style to natural photographs and show that training the same segmentation network on a surrogate artworks *improves the accuracy for fine art paintings. We also provide a dataset with pixel-level annotation of 60 fine art paintings to the public and for evaluation of our method.*

Introduction

Semantic image segmentation is a key step in many high-level tasks in image analysis, including object recognition, captioning, and question-answering applied to natural photographs, lane-following and obstacle avoidance applied to real-time videos in autonomous vehicles, identification and quantification of biological structures such as tumors applied to medical images as well as quantification and analysis of land, water, vegetation, and geographic structures applied to remotely sensed images. Currently, the most accurate approaches to such problems have relied on training deep neural networks with large corpora of images representative of the ultimate task, such as natural photographs, video stills from on-board vehicle cameras, medical x-ray and fMRI scans, and multi-spectral satellite images. [3]

Semantic image segmentation is similarly a key step in many tasks in the analysis of fine art paintings and drawings, such as quantification of composition, style, question answering, and semantic interpretation. [18] Fine art images differ in several important ways from the images mentioned above, however, and for this reason, the prior approaches to segmentation do not work adequately, as reported in our earlier work,[4] and as we shall see in the baseline results. The reasons for such difficulties include:

Wide variety of styles Two-dimensional artworks—even those restricted to representational works of the Western canon—display an immense variety of styles and media, as is evident from artists as diverse as Duccio, Leonardo, Caravaggio, Rembrandt, El Greco, Egon Schiele, Karel Appel, James McNiell Whistler, Gustav Klimt, Winslow Homer, Georges Braque, Alice Neel, Alex Katz, and Richard Estes. The stylistic variations may include lighting, color palette, the nature of marks and brushstrokes, stylistic variations in shapes of contours, and much more.

Small corpora The number of images of two-dimensional fine art is far smaller than the size of corpora of natural photographs used in computational image analysis, which may number in the hundreds of millions or in some private commercial settings over a billion. By contrast typical numbers of known artworks for specific artists are rarely larger than a few thousand and with large variation in styles and content.

Non-physical conventions Many representational artworks exploit non-physical conventions, such as glowing infants, flying putti, the absence of cast shadows, inconsistent geometric perspective, and more.

Imaginary objects Some artworks depict imaginary creatures and objects such as dragons, gryphons, winged angels, multi-headed serpents, enormous parting seas or gods holding thunderbolts amidst the clouds.

Rare scenes or events Many artworks, particularly religious works from the European Renaissance and Baroque, depict scenes that may conform to the laws of physics but are nevertheless extremely rare or unusual, such as crucifixion, a woman decapitating a king, a sinking raft crowded with sailors after a shipwreck, and innumerable other examples.

For these and related reasons, natural photographs taken alone do not provide representative images for the development of algorithms for most tasks in processing and analyzing images of fine art, a point we shall illustrate in our base-case results.

In this work we address these issues by applying style transfer methods to public datasets of natural photographs to create a surrogate dataset in the style of a given artist. Our approach is motivated by recent works on domain adaption with style transfer [1, 2, 6]. These works showed improvements for Monocular depth prediction and semantic segmentation by training on style transferred datasets. However the styles used were based on realistic images. Artworks often depict objects in styles and textures different from how similar objects are pictured in natural images. We hypothesize that training a semantic segmentation network on a surrogate dataset in the style of a specific artist can improve the segmentation results on real artworks in the same style even for very highly stylized and non-realistic works. We

† equal contribution

select four artists with increasingly abstract styles (Egon Schiele, Vincent Van Gogh, Pablo Picasso, Willem de Kooning) to measure the performance improvements and the level of abstractions our method can manage. The unsatisfactory generalization ability of current semantic segmentation methods to styles not seen during training is partly the premise of our work. However, the works of artists such as Pablo Picasso exhibit significant changes in style over the course of their lives. For this reason, we limit our study to a single style or "period" per artist, such as Willem de Kooning's Woman series series.

After an introduction to existing work in the areas of a and domain adaptation, we give an overview of our system. Then, in the training section, we describe our datasets and training protocols. In the results section, we show typical segmentation results for paintings by 4 artists and conclude the paper with a reflection on our results and opportunities for future improvements.

Related Work

Recent methods for semantic segmentation rely on large quantities of labeled data for the target domain. Transfer learning [19] can be employed to improve performance in cases where less labeled data is available. However, in cases where no labeled data for the target domain is available, a different approach is needed. One such method that has been shown to work well in particular for fine art paintings is style transfer [8, 7, 12]. Style transfer describes a technique to transfer the artistic style of a source image *S* to a content image *C*. The result is a new image that maintains the global structure of the content image *C* while having the local structure of the style image *S*. The method was initially presented in Gatys et al. [8]. There, the authors use a VGG 19 network [17] to extract features from style image *S* and content image *C* by extracting the output of the middle layers of the network. A third image, the generated image *G* is then optimized such that it contains the content of *C* in the style of *S*. The task is casts as an optimization problem which can lead to long inference times. A faster method is presented in Johnson et al. [13] where style transfer is performed by a single forward pass of a CNN that is trained with a perceptual loss. However, the method requires training a different network for every style. In Geirhos et al. [9] a style transferred version of imageNet [5] is used as a training set for a CNN image classifier which is shown to reduce texture bias and favours learning shape-based representations.

System overview

As in the earlier work, our current overall system architecture and methodology relies on two deep networks, a STYLE TRANSFER network and a SEMANTIC SEGMENTATION network.

Each module is trained differently and has a different function, as we now describe.

Style transfer module

Style transfer is performed with the method presented in Gatys et al. (2016) [8]. There, a VGG 19 network, trained for image classification on imageNet, is applied to both a content image *C* as well as a style image *S*. The latent embeddings after 5 different layers in the network are then extracted as feature vectors. A third image *G*, that we initialize with the content image, is then optimized such that its content matches *C* and its style matches *S*. We optimize the image *G* to match the content of *C* and the style of *S*. The loss is the sum of the content loss (mean squared error between features of *G* and features of *C*) and the style loss that measures the correlation between the feature channels of *S* and the feature channels of *G*. The cross-connections mean that the final output image represents an integration of the two and leads to the source image rendered in the style of the style image.[8] The cross-connections mean that the final output image represents an integration of the two, and leads to the source image rendered in the style of the style image.[8] Figure 1 shows three representative natural photographs and three artworks from our database and the resulting transfer of style. The three images in the bottom row are thus surrogate artworks, each member of a separate training set for the segmentations of works by the associated target artist.

While the surrogate works certainly appear subjectively to bear the styles of the associated target artists, we do not have a principled way to quantify the perceptual quality of the style transfer —nor do we need such a measure. After all, our central task is semantic segmentation, which does have a principled measure of performance. Surely the performance as gauged by this measure depends upon both the quality and quantity of the surrogate artworks.

Semantic segmentation module

The semantic segmentation module is a Convolutional Neural Network (CNN) based on the Pyramid Attention Network architecture from Li et al. [14]. In the case of Vincent Van Gogh we initialize the network with a ResNet34 [10] backbone while a ResNet50 backbone is used for the others. We use two output classes for binary segmentation and 7 for multi-class segmentation (Van Gogh landscape paintings).

Training

Our method is trained and evaluated for two application cases:

1. Binary segmentation of Persons (Styles from artists Pablo Picasso, Egon Schiele and Willem de Kooning)
2. Multi-class segmentation with the classes ground, sky, water, building, plant and person (Style from paintings of Vincent Van Gogh)

We create ground truth segmentation labels for 58 paintings (17 Picasso paintings from the African period, 11 Schiele portraits, 24 van Gogh paintings of landscapes and 6 de Kooning paintings from his *Woman* series) for testing and as style images. The number of works per artist differ, depending on the availability of artworks for the specific style. Our method is trained on stylized versions of the Microsoft COCO dataset [15]. For the binary segmentation networks applied to Pablo Picasso, Egon Schiele and Willem de Kooning we extract a subset of 7931 images split into 7593 training images and 338 validation images. In these cases, the COCO dataset is filtered for images semantically similar to the studied portrait artworks. In particular, images are required to contain at least one person, and the area of all persons has to be at least 30% of the image. For the more complex multi-class segmentation network applied to the works of Vincent Van Gogh with the classes ground, sky, water, building, plant and person, we extract a subset of 46348 images split into 43297 training and 3051 validation images. For the artists Pablo Picasso, Egon Schiele and

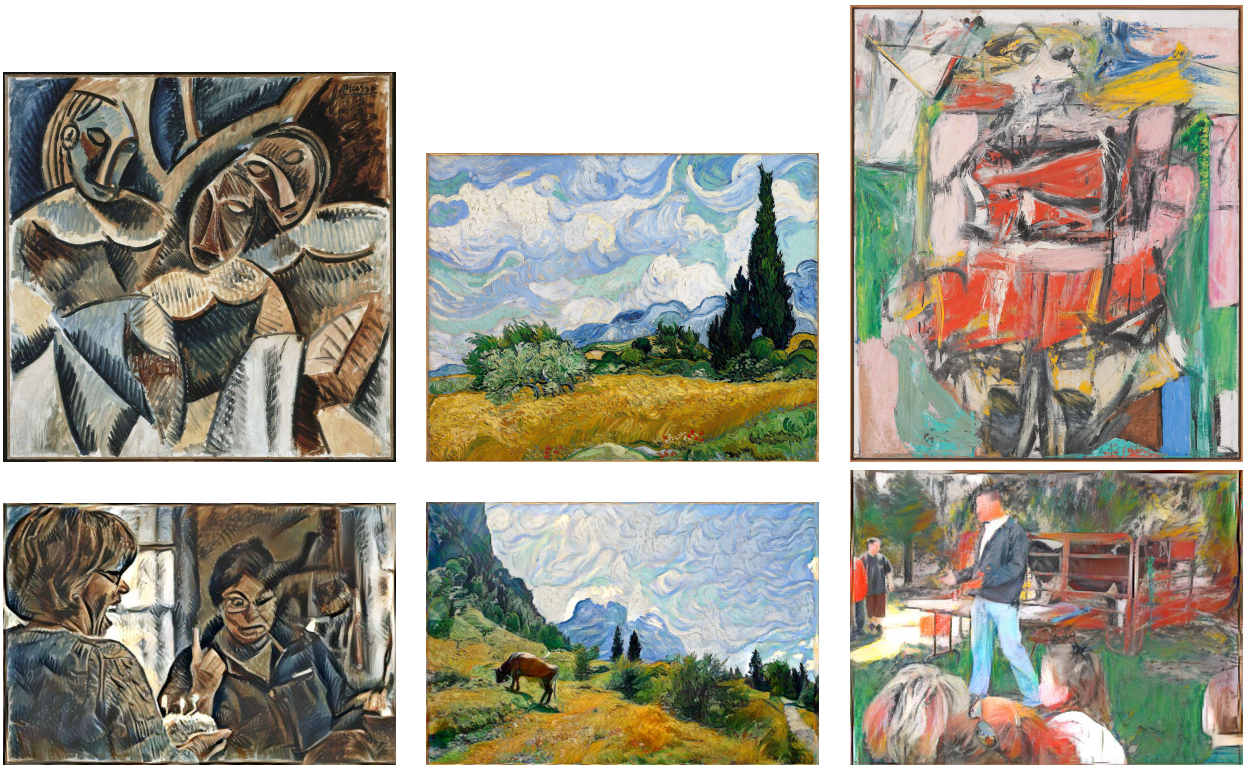

Figure 1. *Top row: representative artworks by three artists, Pablo Picasso's* trois figures sous un arbre picasso *(1907-1908), Vincent Van Gogh's* Wheat Field with Cypresses *(1889), and Willem de Kooning's* Woman VI *(1953). Bottom row: natural photographs rendered in the style learned transferred from the artist above.*

Willem de Kooning we create stylized versions of the binary segmentation subset and for Vincent Van Gogh we create a stylized version of the multi-class segmentation subset. To better capture the background style for Schiele images, we apply the GT mask to the training images and set pixels outside the mask to white (Both for the baseline run and for the run on style transferred images). In all cases, the networks are trained to minimum error on the validation set.

For an image of height N and width M we define its domain $D := \{1 \,..\, N\} \times \{1 \,..\, M\}$. Its corresponding ground truth mask $G = (G_{ij})_{(i,j)\in D}$ and predicted mask $P = (P_{ij})_{(i,j)\in D}$ are assumed to have values in the range $[0,1]$. For training we find that the combination of the three loss functions mean squared error $\mathcal{L}_{\text{MSE}}$, dice loss $\mathcal{L}_{\text{DSC}}$ [16] and the gradient loss $\mathcal{L}_{\nabla}$ [11] is best suited for our segmentation task:

$$\mathcal{L}_{\text{MSE}}(P,G) := \frac{1}{NM} \sum_{(i,j)\in D} \big(P_{ij} - G_{ij}\big)^2, \tag{1}$$

$$\mathcal{L}_{\text{DSC}}(P,G) := 1 - \frac{\varepsilon + 2\sum_{(i,j)\in D} P_{ij} G_{ij}}{\varepsilon + \sum_{(i,j)\in D} P_{ij} + G_{ij}}, \tag{2}$$

$$\mathcal{L}_{\nabla}(P,G) := \mathcal{L}_{\text{MSE}}(\nabla P, \nabla G). \tag{3}$$

Here $\varepsilon \ll NM$ is a small regularization constant that we set to 10^{-5} for our experiments. The loss function is then given as the sum of these: $\mathcal{L} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{DSC}} + \mathcal{L}_{\nabla}$. The gradient loss enables a better separation between different semantic regions due to its focus on changes in the ground truth mask, while the dice loss is closely related to the IoU metric and provides a more holistic loss signal compared to per-pixel distance measures such as mean squared error. Compared to mean squared error and cross entropy loss functions, we observe an up to 15% improvement in segmentation quality as measured by the IoU metric. In the case of multi-class segmentation with N the loss signal is formed as a sum over all classes, i.e. $\mathcal{L}_N(P,G) = \sum_{c=1}^{N} \mathcal{L}(P_c, G_c)$, where P_c and G_c are binary predictions and segmentation masks.

For all of our experiments, we use the Adam optimizer with an initial learning rate of 10^{-3} and reduce it after 15 and 20 epochs by a factor of 10. The available compute infrastructure consists of an Intel Xeon E5-2697v3 CPU (14 cores), 128GB RAM and 4 Nvidia GeForce GTX Titan X GPUs. The entire setup is implemented in PyTorch 1.9.0.

Results

We now turn to the segmentation results, first the baseline segmentations using a network trained on only natural photographs, and then with a network trained with the artist-specific surrogate images.

Our measure of quality is the mean intersection over union (mIoU) metric defined as

$$\text{mIoU}(P,G) = \frac{1}{N} \sum_{c=1}^{N} \frac{|P_c \cap G_c|}{|P_c \cup G_c|}, \tag{4}$$

where N is the number of classes, $0 \le \text{mIoU} \le 1$, and the higher the score the better. Before evaluation we apply a threshold of 0.5 to obtain binary predictions.

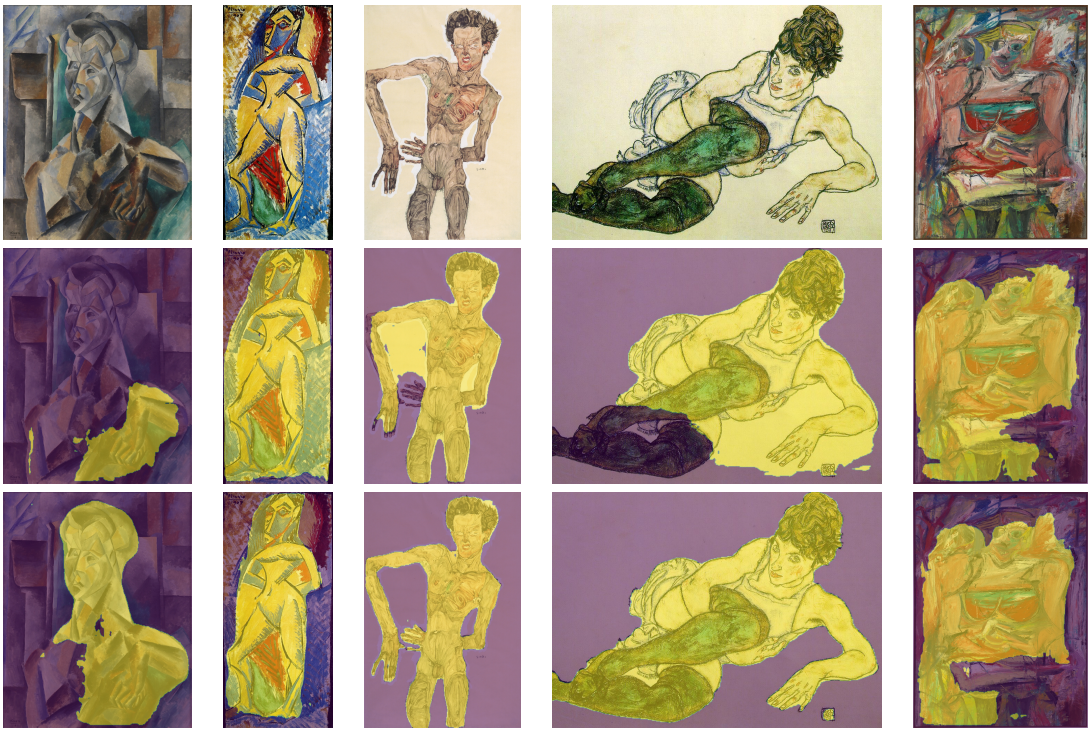

Figure 2. *Top row: artworks. Middle row: Baseline segmentation using deep net trained on only natural photographs. Bottom row: Segmentation based on network trained using artist-specific style transferred images. a) Pablo Picasso's* Buste de femme*, b) Pablo Picasso's* Femme Nue*, c) Egon Schiele's* Nude Self-Portrait, Grimacing*, d) Egon Schiele's* Reclining Woman with Green Stockings*, e) Willem de Kooning's* Woman I.

Baseline results

The top row in Fig. 2 shows works by the four artists we tested for the task of binary person segmentation. The middle row shows the baseline segmentation results. Notice the errors of omission and commission. These indicate that the neural network has difficulty recognizing the separating boundary between foreground and background - an issue that is especially apparent in the work of Pablo Picasso and taken to the extreme with Willem de Kooning's Woman series. In such cases even humans are unable to select a definite boundary between semantic regions. Fig. 3 shows the predicted segmentation for Vincent van Gogh's White House at Night. The baseline results are depicted in the top row. The IoU scores are shown in Table 1.

Table 1: Segmentation results where n denotes the number of artworks tested. For Picasso, Schiele and de Kooning we beform binary segmentation and for van Gogh semantic segmentation with 6 classes.

Artist	n	Baseline	Final	Improvement (%)
Picasso	17	38.70	54.60	41%
Schiele	11	72.70	82.22	13%
de Kooning	6	63.81	47.93	–
van Gogh	24	22.83	34.47	51%

Style-transferred results

The bottom row in Fig. 2 shows representative results for the three binary segmentation tasks using the network trained with surrogate artworks based on the styles in each of the target artists. As can be seen in column a, where the results for Pablo Picasso's Buste de femme are depicted, training on style transfered images improved the detection of the human shape and the seperation of the figure from its background. The example in column b shows an improvement in false positive errors, however the produced segmentation mask remains a super-set of the ground truth. The works of Egon Schiele (columns c and d) show remarkable improvements in segmentation quality with close to perfect results in many cases. Our method fails to improve results for Willem de Koonings's Woman series, as can be seen in the example in column e where Willem de Kooning's Woman I is depicted. Fig. 3 depicts the multi-class segmentation results for Vincent van Gogh's White House at Night when trained on the style-transferred dataset. Our method leads to improved results for all six classes. Table 1 shows the IoU scores. We find that our method improves the segmentation for all artists except Willem de Kooning.

Finally, Table 2 shows the performance of models trained on a dataset with the style given by the row name evaluated on samples from the artist given by the column name. We find that, except for Willem de Kooning, where our method fails to improve the segmentation results, the segmentation results are best when training the method on stylized data with the same style as the test set.

Figure 3. *Top row: Baseline segmentation using deep net trained on only natural photographs. Bottom row: Segmentation based on network trained using artist-specific style transferred images.*

Table 2: Segmentation results comparison between styles. The row names represent the style of the training dataset and the column names show the artist of the test artworks.

	Artist		
Network	Picasso	Schiele	de Kooning
Picasso	**54.60**	68.86	**53.00**
Schiele	51.57	**82.22**	37.16
de Kooning	50.96	63.12	47.93

Conclusions and future directions

We have demonstrated a method for generation of surrogate images in the style of particular artists, and that training semantic segmentation networks with such datasets leads to significant improvement of fine art images by target artists. We created ground truth segmentation labels for artworks from four different artists.

Our general technique of creating large corpora of surrogate art images should find use in several tasks of art analysis beyond segmentation. For instance, object recognition and derivative functions (such as counting, localizing, and so on), and captioning. Such tasks, when applied to natural photographs, require very large corpora of training data.

With our method, we identify two causes for incorrectly segmented image regions, which can independently serve as a basis for future work. The first is image regions that – to the human eye – clearly belong to one of the predefined semantic classes, but were misclassified due to the lack of generalization capability of the segmentation network. Errors of this type can be reduced by more training data or improvements in the network architecture. The second type of error results from altered content in the surrogate artworks. If the style component used for neural style transfer is too abstract, the algorithm may incorrectly apply the style of one semantic category to another (e.g., applying the style of trees in the background to a person's face). It is in this second area, and the computational requirements of neural style transfer, that we see the greatest potential for improvement.

Specifically, we propose two major paths:

1. Improving domain adaption for very abstract styles. Our method failed to improve the performance for paintings from Willem de Kooning from the *Woman* series which is most likely a consequence of the highly abstract nature of the series together with the fuzziness of the edges between person and background therein. The style transfered result for Kooning in Fig. 1 shows that style transfer fails to deform people in a similar way as in Koonings portrait series.
2. Reducing the time needed for style transfer by switching to a feed-forward method similar to [13] would allow applying the style transfer as a form of data augmentation during training.

Acknowledgements

This work stems from a student project in *Computer vision and image analysis of art* offered by the last author at the Technical University of Vienna January–February, 2021. We thank Professor Robert Sablatnig for support of both that class and this current extension of the original project. The last author would like to thank the Getty Research Center, Los Angeles, in particular its Research Library, where some of this work was performed.

References

[1] Paolo Andreini, Simone Bonechi, Monica Bianchini, Alessandro Mecocci, and Franco Scarselli. Image generation by gan and style transfer for agar plate image segmentation. *Computer methods and programs in biomedicine*, 184:105268, 2020.

[2] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2810, 2018.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Applications*, 39(12):2481–2495, 2017.

[4] Anthony Bourached, George H. Cann, Ryan-Rhys Griffiths, and David G. Stork. Recovery of underdrawings and ghost-paintings via style transfer by deep convolutional neural networks: A digital tool for art scholars. In David G. Stork and Kurt Heumiller, editors, *Computer vision and analysis of art*, pages xxxx–xxxx. IS&T, Springfield, VA, 2021.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Aysegul Dundar, Ming-Yu Liu, Ting-Chun Wang, John Zedlewski, and Jan Kautz. Domain stylization: A strong, simple baseline

for synthetic to real image domain adaptation. *arXiv preprint arXiv:1807.09384*, 2018.

[7] LA Gatys, AS Ecker, M Bethge, A Hertzmann, and B Shechtman. Controlling perceptual factors in neural style transfer. In *30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 3730–3738. IEEE, 2017.

[8] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, New York, NY, 2016. IEEE.

[9] 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 *International Conference on Learning Representations*, 2018.

[10] Kaiming He, Xiangyu Zhang, Shoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Thomas Heitzinger and Martin Kampel. Highly accurate binary image segmentation for cars. *ARW & OAGM Workshop*, 2020.

[12] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019.

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[14] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[18] David G. Stork. Automatic extraction of meaning in authored images such as artworks: A grand challenge for AI. *ACM Journal of Computing in Cultural History Computing*, xx(xx):xx–xx, 2021a.

[19] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Author Biography

Matthias Wödlinger received his MSc in physics and MSc in mathematics from the Technical University of Vienna in 2018. Since then he has worked at the Computer Vision Lab at the Technical University of Vienna. His work has focused on applications of deep learning in document analysis, cancer research and image compression as well as deep learning theory.

Thomas Heitzinger holds an MSc in Applied Mathematics and an MSc in Logic and Computation. He is currently a PhD student at the Computer Vision Lab (CVL) at TU Wien. His main research interests are semantic segmentation for sepcialized target applications and 3D scene understanding on resource constrained hardware and human behavior analysis using depth sensors and other non-RGB-based visual sensors, as well as computer vision without strong texture information.

David G. Stork holds degrees in physics from the Massachusetts Institute of Technology and the University of Maryland at College Park. He has made technical contributions in theoretical mechanics, computational imaging, computer vision, optics, machine learning, pattern classification, computational data acquisition, concurrency theory, cryptography, visual psychophysics and perception, statistics, combinatorics, and other areas. He studied art history at Wellesley College and was an artist-in-residence at the New York State Council of the Arts.