The 8th ACM/EG Expressive Symposium
C. Kaplan, A. Forbes, and S. DiVerdi (Editors)

Real-Time Patch-Based Stylization of Portraits Using Generative Adversarial Network

D. Futschik[1], M. Chai[2], C. Cao[2], C. Ma[2], A. Stoliar[2], S. Korolev[2], S. Tulyakov[2], M. Kučera[1], D. Sýkora[1]

[1]Czech Technical University in Prague, Faculty of Electrical Engineering, Czech Republic
[2]Snap Inc., USA

Figure 1: *Given an input exemplar and a target portrait photo, we can generate stylized output with comparable or superior visual quality as compared to several state-of-the-art face stylization methods (Fišer et al. [FJS*17], Liao et al. [LYY*17], Selim et al. [SED16], and Gatys et al. [GEB16]) while being able to run at interactive frame rates on a consumer GPU. Style exemplar:* © Scary Zara Mary.

Abstract

*We present a learning-based style transfer algorithm for human portraits which significantly outperforms current state-of-the-art in computational overhead while still maintaining comparable visual quality. We show how to design a conditional generative adversarial network capable to reproduce the output of Fišer et al.'s patch-based method [FJS*17] that is slow to compute but can deliver state-of-the-art visual quality. Since the resulting end-to-end network can be evaluated quickly on current consumer GPUs, our solution enables first real-time high-quality style transfer to facial videos that runs at interactive frame rates. Moreover, in cases when the original algorithmic approach of Fišer et al. fails our network can provide a more visually pleasing result thanks to generalization. We demonstrate the practical utility of our approach on a variety of different styles and target subjects.*

CCS Concepts
• ***Computing methodologies*** → *Non-photorealistic rendering;*

1. Introduction

The stylization of human portraits becomes highly attractive thanks to the massive popularity of selfie photography and invention of mobile applications such as MSQRD or Snapchat which use facial landmarks together with CG rendering pipeline to deliver stylized look. This approach, however, requires professional artists to carefully design textured 3D models along with custom shaders to achieve the desired look.

This limitation can be alleviated using example-based approaches pioneered by Hertzmann et al. [HJO*01]. This technique allows transferring style from a given artistic exemplary image to a target photo. State-of-the-art in this domain uses neural-based techniques [SED16], patch-based synthesis [FJS*17], and their combinations [LYY*17] to deliver impressive stylization results. However, a key limitation of those techniques is that they consist of several algoritmic steps each of which may be a source of potential failure (see Figures 5, 6, and 7, two right columns) and introduces algoritmic complexity which leads to huge computational overhead.

Generative adversarial networks [GPAM*14] have become a favorite technique for image-to-image translation tasks [IZZE17, WLZ*18a, WXWT18] recently. Their principal drawback over

classical style transfer techniques which require only a single style exemplar image [GEB16] is the necessity of training the network on a large dataset of paired appearance exemplars. This requirement is prohibitive in the case of artistic style transfer as tedious manual work is necessary to prepare the training dataset. Although unpaired alternatives exist [ZPIE17, ZZP*17] they still require many drawings of a particular style as an input. Another issue is related to the fact that current image-to-image network architectures have difficuties in reproducing delicate high-freuqency details that are important to retain fidelity of used artistc media.

In this paper, we demonstrate the benefits of combining state-of-the-art high-quality patch-based synthesis with the power of image-to-image translation networks. Thanks to the ability of patch-based method of Fišer et al. [FJS*17] to produce high-quality results we can generate a dataset which preserves the original artistic style precisely. We then use this dataset to train a variant of image-to-image translation network with improved structure that better preserves important high-frequency details. Although the method of Fišer et al. is prone to failure in more complex cases, we laverage the fact that the network can generalize even when the training dataset contains many failure exemplars. This beavhior was recently demonstrated in a different context of generative models trained from partially observed samples [BPD18] or without ground truth counterparts [LMH*18]. Thanks to this ability to generalize while still being able to preserve high-frequency details, we can produce results which are comparable or sometimes more visually pleasing than the output of the original patch-based method. Moreover, since the trained network can be evaluated quickly on the GPU our approach enables real-time style transfer which was unattainable for previous high-quality techniques.

2. Related Work

The stylization of head portraits is a long-standing challenge for non-photorealistic rendering (NPR) research community. In this domain, traditional filtering-based stylization techniques [GRG04, TL05, DiP07, YLL*10] have been extensively used to deliver compelling results for simple styles. However, they do not allow for greater appearance variations.

Example-based techniques can be used to alleviate this limitation. One possible solution is to compose the final image using a set of stylized facial components prepared by an artist [CLX*02, CZL*02, CLR*04, MZZ10, ZDD*14]. Although this approach provides greater freedom for local regions, it is still challenging to preserve the identity of the target person due to the inability to adapt the templates to the unique geometry of target facial features.

To overcome this drawback, researchers further propose to prepare a larger dataset of photo-style exemplary pairs (e.g., *CUHK Face Sketch Database* [WT09]), and then use multi-scale Markov Random Fields [WT09, LLN11, ZKW12, WTG*13, WTG*14] to estimate the stylization for a given target face. Although these techniques can deliver better identity adaption, they are highly impractical since many photo-style exemplars need to be prepared manually for each new artistic style.

The example-based approach can also be reduced to the level of individual brush strokes [ZZ11, BSM*13, WCHG13]. Although these techniques are compelling at delivering particular artistic looks (e.g., oil paint), they are difficult to apply on styles where the interaction between individual brush strokes cannot be modeled merely by blending operations.

Recently, neural network based style transfer becomes very popular thanks to the seminal work of Gatys et al. [GEB16]. The success of this method motivated others [SED16, LZY*17] to develop custom neural-based stylization techniques for human portraits. Although those example-based methods can achieve generally compelling results, they usually fail on more complex structured exemplars where preserving high-frequency details is critical. Recently, patch-based techniques [FJS*17, LXZ*18] have been proposed that try to address this issue. Nevertheless, these require additional guiding channels to be prepared, which govern the synthesis process to transfer patches in a semantically meaningful way between the style exemplary and the target photo. Although such channels can be created automatically via a series of algorithmic detectors, this solution makes the system more fragile as an occasional failure of any individual unit may significantly affect the whole synthesis.

Li et al. [LW16] introduce a combination of neural- and patch-based synthesis. Their key idea is to use responses of a deep neural network trained on image classification [SZ14] to establish patch-wise correspondences between the style exemplar and the target image. Liao et al. [LYY*17] and Gu et al. [GCLY18] later extended this approach to perform patch-based synthesis directly in the domain of latent neural feature spaces, and then reconstruct the final image using deconvolution. Recently, Cao et al. [CLY18] propose to perform geometric exaggeration on top of appearance transfer. Despite the impressive results, these techniques still suffer from common pixel-level artifacts which lead to lower quality of the synthesized imagery as compared to patch-based methods which can work directly in the image domain and preserve important pixel-level details.

Our approach bears a resemblance to techniques which can quickly perform certain image editing operations for which time consuming algorithmic solutions exists [XRY*15, CXK17]. By training a feed-forward network on a pre-computed dataset they can achieve significant speed up as well as a level of generalization that sometimes outperforms quality of results produced by the original algorithm. A similar technique was also used in the context of neural-based style transfer by Johnson et al. [JAFF16]. In their approach, the output from Gatys et al.'s algorithm [GEB16] was used to train the weights of a feed-forward neural network. However, as Gatys et al.'s method does not perform semantically meaningful transfer the ability to generalize and increase the robustness was not as apparent.

The tendency to generalize and improve upon the original training dataset has been recently reported also in the case where corrupted datasets are used for training [BPD18, LMH*18]. In these works authors observed the ability of a generative network to recover from failures and produce comparable or sometimes even better visual quality as compared to a scenario when a clean dataset is used for training.

Recently, there were attempts to generalize neural-based stylization [LFY*17, HB17] so that costly training nor optimization is required to perform fast style transfer from arbitrary exemplar.

Nevertheless, those techniques are unable to perform semantically meaningful transfer and still suffer from visible pixel-level artifacts which decrease their ability to reproduce important visual characteristics of used artistic media.

3. Our Approach

Our goal is to learn a mapping function F between color images of human faces $\mathbb{X}$, and their stylized counterparts $\mathbb{Y}$. Since in our case paired data can be produced easily using the algorithm of Fišer et al. [FJS*17], we can model the mapping as a direct transformation $F : \mathbb{X} \rightarrow \mathbb{Y}$.

Given pairs of training samples: $(x_i, y_i)_{i=1}^{N}$ where $x_i \in \mathbb{X}$ and $y_i \in \mathbb{Y}$, our objective to learn F contains three different terms: *adversarial loss* $\mathcal{L}_{GAN}$ for matching the distribution of generated images to the distribution of the stylized images [GPAM*14], a *color loss* calculated directly on the stylized output $\mathcal{L}_1$, and finally a *perceptual loss* $\mathcal{L}_{VGG}$ calculated on features extracted by a VGG network pre-trained on ImageNet [SZ14]. In the following section we focus on each loss in more detail and state the final objective function. Then we describe our network architecture and dicsuss implementations details.

3.1. Training Objective

Adversarial Loss. We apply adversarial loss to the output of the mapping function F and its discriminator D_Y using the following objective function:

$$\mathcal{L}_{GAN}(F, D_Y, \mathbb{X}, \mathbb{Y}) = \mathbb{E}_{y \sim p_{data}(y)}\left[\left(D_Y(y) - 1\right)^2\right] + \mathbb{E}_{x \sim p_{data}(x)}\left[\left(D_Y(F(x))^2\right)\right] \quad (1)$$

where instead of traditional binary cross entropy $\mathcal{L}_2$ norm is used as the adversarial criterion. This leads to a more stable training [MLX*17].

Color Loss. While adversarial loss alone could be enough to learn mapping F, we observed that when an additional $\mathcal{L}_1$ loss [IZZE17] is computed between the output of the network and the original stylized image we can encourage the generator to better preserve identity as well as stabilize and speed up the training:

$$\mathcal{L}_1(F) = \mathbb{E}_{X,Y \sim p(X,Y)} ||Y - F(X)||_1 \quad (2)$$

Perceptual Loss. Additional improvement can be achieved using perceptual loss that is calculated on feature maps of the VGG-19 model pre-trained on ImageNet at different depths:

$$\mathcal{L}_{VGG}(F) = \sum_{d \in D} ||VGG_d(Y) - VGG_d(F(X))||_2 \quad (3)$$

where D is the set of depths of VGG-19 which are considered, in our case $D = 0, 3, 5, 10$. Similar approach was used also in [WLZ*18a], however, Wang et al. used $\mathcal{L}_1$ norm which we found has notably lower impact on the final visual quality as compared to our $\mathcal{L}_2$ norm (see Figures 2a and 2c).

Objective. Using all mentioned losses our final objective function is as follows:

$$\mathcal{L}(F, D_Y, X, Y) = \lambda_1 \mathcal{L}_{GAN} + \lambda_2 \mathcal{L}_1 + \lambda_3 \mathcal{L}_{VGG} \quad (4)$$

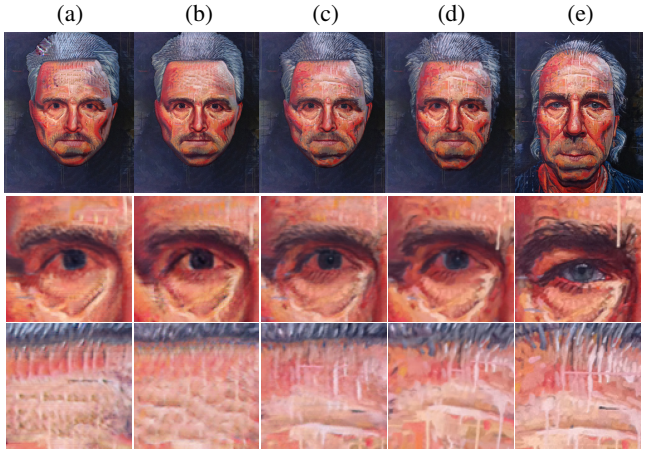

Figure 2: *Ablation study. A demonstration of visual quality improvement achieved using modified VGG loss and our improved network architecture: (a) result of our network trained without using VGG loss, (b) result generated using all losses, however, without our improved network architecture, i.e., using the original architecture of Johnson et al. [JAFF16], (c) our result, (d) result generated using FaceStyle algorithm [FJS*17], (e) style exemplar. Note how our full-fledged approach better reproduces the original style exemplar (see the avoidance of artificial repetitive patterns on forehead as well as sharper details around eyes) and also slightly improve upon the output of FaceStyle algorithm (c.f. better preservation of important facial features like ears or nose). Style exemplar:* © Matthew Cherry *via http://matthewivancherry.com/home.html and https://www.instagram.com/matthewivancherry.artist (HAT, oil on canvas, 48" x 48", 2011).*

where $\lambda_1, \lambda_2, \lambda_3$ influence the relative importance of the different loss functions.

3.2. Network Architecture

For our generator model we use the initial architecture from [JAFF16], three convolution blocks (two of them with stride = 2) which are followed by several residual blocks [HZRS16], two upsampling blocks and finally a tanh activation. Compared with Johnson et al.'s solution, we make the following modifications (see Fig. 3) which we observed had a significant impact on the final perceptual quality: we changed the size of convolutional filters in the very first layer from 9×9 to 7×7 and in the very last layer of the original architecture from 9×9 to 5×5. We increased the number of residual blocks used from five to nine. Next, we added skip connections using concatenation of feature maps [RFB15] to the upsampling layers, which has been shown to improve gradient propagation, and we replace convolutions with fractional strides with nearest neighbor upsampling followed by an additional 3×3 convolution. Lastly, we attached two more convolutional layers before the output, which we observed have positive effect when the skip connections are added. All these modifications helped to preserve important high-frequency details in the generated image (see

visual quality improvement over the initial generator's structure in Figures 2b and 2c).

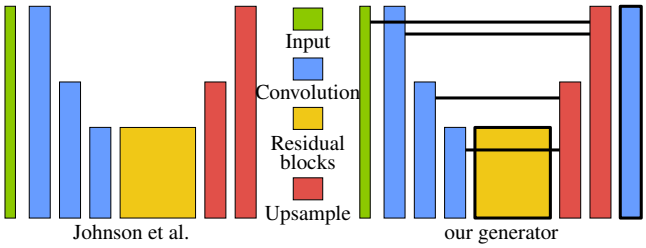

Figure 3: *The original generator network architecture of Johnson et al. [JAFF16] (left) followed by our improved architecture (right). Modifications are denoted with black color: added skip connections, increased the number of residual blocks, two upsampling layers are followed by additional transposed convolution layer.*

For our discriminator model we use PatchGAN model [IZZE17] using progressively higher number of feature maps with instance normalization proposed by Ulyanov et al. [UVL16] and leaky ReLUs as activation. This helped us to lower the number of parameters and achieve a more stable gradient propagation.

3.3. Implementation Details

We implemented our approach using C++ and the Python framework PyTorch.

For FaceStyle algorithm we used settings recommended in the original paper [FJS*17]. For each artistic style we produced a training set of 5124 stylized facial images in a resolution of 512×512 which is supported by our network architecture. We used automatic portrait segmentation [SHJ*16] to assure the training algorithm focus more on important facial parts of the input image. Since we did not pre-filter the dataset the resulting set of samples contains both successful as well as failure exemplars (c.f. two right columns in Figures 5, 6, and 7 to see examples of such failures).

For training of our models we used the Adam solver [KB14] with a batch size of 2. In total, our generator model has 14.7 million parameters, and our discriminator has total number of parameters of 694 thousand. We set $\lambda_1 = 0.3$, $\lambda_2 = 5$, and $\lambda_3 = 0.7$, which were chosen experimentally via grid search and manual tuning. Both generator and discriminator networks were trained from scratch with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $lr = 0.0002$. During the training phase we found that we could use as few as 2000 samples without significant loss of quality. Sufficiency of lower number of training samples can be explained by limited complexity of the appearance changes in the stylized output. We train our models in 50 epochs. Some styles proved to be more challenging to learn, and thus we allowed training in 100 epochs. In general, training for one epoch took around 83 minutes on a single NVIDIA Tesla P100 GPU, making the total training time for one style slightly shorter than 3 days.

4. Results

We trained our network on seven different style exemplars (see Figures 1, 2, 4 and 9) and applied it to 24 portraits not included in the training dataset. In Figures 1, 2, 5, 6, 7, and 9 results of our trained network are compared with the original FaceStyle algorithm [FJS*17].

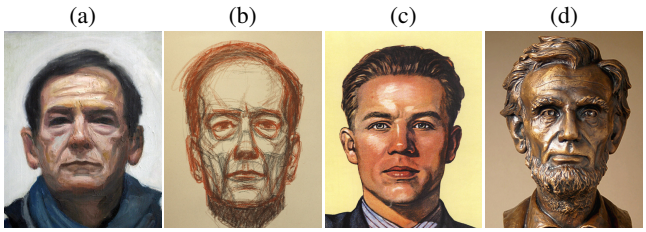

Figure 4: *Exemplars of styles used in Figures 6, 7, and 8. See Figures 1, 2, and 9 for the remaining style exemplars. Style exemplars: (a–b)* © Adrian Morgan*, (c)* Viktor Ivanovich Govorkov*, (d)* © Will Murray.

In the following sections, we discuss potential of our method to perform real-time high-quality style transfer, we also mention its ability to generalize and increase robustness over the original FaceStyle algorithm [FJS*17] and describe a perceptual study we conducted to evaluate visual quality of our approach with respect to the output generated by FaceStyle algorithm. Finally, we compare our results with current state-of-the-art.

4.1. Interactive Scenario

Thanks to the compactness of our network (47MB) we can perform feed-forward propagation in real-time (15 frames per second) on currently available consumer graphics cards (we use GeForce RTX 2080 Ti). This benefit enables us to implement the first high-quality style transfer on live video streams (please refer to our supplementary video). We can downsize our network architecture to 256×256 resolution (along with reducing the number of filters in each layer) and also achieve interactive response on mobile devices without significant loss of visual quality.

4.2. Generalization

During the training experiments we found that when we deliberately filter out failure exemplars from the training dataset the overall visual quality does not increase significantly, however, the robustness of the resulting trained network decreases. This behavior bears resemblance to findings reported by Lehtinen et al. [LMH*18] although in our case the nature of corruption cannot be modelled by zero-mean noise, we can characterize this tendency as a convergence to an equilibrium which expresses a "mean" of stylized appearance. Thanks to this behavior the trained network can in practice repair failures of the original FaceStyle algorithm. In cases when the FaceStyle algorithm produces correct result our network can deliver stylization which is comparable or sometimes even more visually pleasing and better preserving the identity of a stylized person (see Figures 1, 5, 6, 7, 2, and 9).

Another important aspect of the equilibrium mentioned above is that it helps to preserve coherent stylization when the target image does not change considerably. This tendency is essential for

Figure 5: *Face stylization results. In each group of three images, from left to right, we show the input image, our stylization result, and the output from FaceStyle [FJS*17]. The corresponding style exemplars are visible in Figures 1 and 2.*

achieving temporal coherency. In contrast to FaceStyle algorithm or other video stylization techniques [CLY*17, RDB18] that would require explicit treatment of consistency between adjacent frames our technique handles temporal coherency implicitly (see accompanying video demo).

4.3. Perceptual Study

To confirm the quality of results produced by our approach are comparable to those produced by the original FaceStyle algorithm [FJS*17] we conducted a perceptual study. The study had the form of an online questionnaire, where we showed each user the input face, input style, and the output. We asked the user to rate the output in two categories: how well does the stylization preserve the identity of the stylized person, and how well does the stylization reproduce the input style. The ratings were from 1 to 10, 1 being the worst and 10 being the best. The questionnaire featured 6 sets of input images and their outputs for both of the tested methods, making a total of 12 image sets showed to users, which were all being rated in the 2 categories. We deliberately selected results which are comparable with no obvious failures. During the time the questionnaire was open, we have collected 194 responses.

We started with the null hypothesis that there is no statistically significant difference between the quality of the outputs of both tested methods, which we tried to reject based on the collected data using the Student's t-test. In the question of identity preservation, we can reject the null hypothesis with a probability of only 49%, which means there is no statistically significant difference between the scores in this category. Our approach scored an average of 6.76 points and FaceStyle scored an average of 6.87 points, which totals to approximately 1% difference on the 1 to 10 scale, supporting the conclusion of both methods being on par with each other. In regard to the style reproduction category, using the same procedure we can reject the null hypothesis with a probability of 63%, which once again does not represent a significant statistical difference. Our approach scored an average of 8.28 points and FaceStyle scored an average of 8.55 points, making only 3% difference. From these results, we can conclude that the outputs of our approach are on par with the outputs of FaceStyle with only minor differences in the overall quality.

4.4. Comparisons

We compared the visual quality of our approach with current state-of-the-art in image-to-image translation (see Fig. 8). For training we used the same dataset as for our method and tweak the parameters to get as close as possible to the appearance of the origi-

Figure 6: *Face stylization results (continued). In each group of three images, from left to right, we show the input image, our stylization result, and the output from FaceStyle [FJS*17]. The corresponding style exemplars are visible in Figure 4.*

nal style exemplar. Results produced by *pix2pix* method [IZZE17] bear a resemblance to our output concerning the ability to preserve the target person's identity. Nevertheless, the network produces several high-frequency artifacts which affect texture details of the original style exemplar. A part of the problem is caused by the fact that the pix2pix network supports only lower resolution (256×256), however, more importantly, the structure of pix2pix generator tends to introduce various uncanny high-frequency patterns. This issue becomes even more apparent in the case of *pix2pixHD* [WLZ*18a] which can support 512×512 resolution, nevertheless, at high frequencies, it still contains disturbing repetitive patterns which are not present in the original style exemplar. The *starGAN* method [CCK*18] roughly preserves basic facial structure, but it also introduces disturbing high-frequency patterns on top of various low-frequency anomalies which give rise to soft color transitions that are not visible in the original style exemplar.

We also compared our approach with concurrent neural-based techniques that do not require training (see Figures 1 and 9). From the comparison it is apparent that the generic neural-based technique of Gatys et al. [GEB16] has difficulty in preserving semantically meaningful transfer. Selim et al. [SED16] provide an improvement over Gatys et al., nevertheless, they still suffer from a loss of critical visual details. Deep image analogies [LYY*17] produce compelling results concerning visual details, but they often fail to keep the consistency of high-level features which affect the identity of the target subject.

5. Limitations and Future work

We demonstrate that our approach brings comparable or even better visual quality within significantly lower computational overhead when compared to the current state-of-the-art. However, there are still some limitations that can encourage future work.

One of the critical challenges is the accuracy and smoothness of head and hair segmentation masks. Although our method often outperforms FaceStyle algorithm concerning the quality of separation of head and hair segments, in general (especially) the outer hair boundary has some issues with smoothness and shape details (see Figures 5, 6, and 7). One can mitigate this inaccuracy by preparing a broader set of training exemplars containing a greater variety of input photos under different illumination conditions with more accurately specified head and face masks.

For some styles our method tends to produce repetition artifacts visible principally on hair segments depending on the overall spatial extent (see Figures 5, 6, and 7). Although a similar effect is

Figure 7: *Face stylization results (continued). In each group of three images, from left to right, we show the input image, our stylization result, and the output from FaceStyle [FJS*17]. The corresponding style exemplars are visible in Figures 4 and 9.*

apparent also on the original output from the FaceStyle algorithm, our solution tends to exaggerate it. Techniques to reduce visible repetition on the level of patch-based synthesis as well as during the training phase (e.g., using a specific penalizing loss) would be a promising avenue for future work.

When inspecting results closely on a pixel level (see Figures 5, 6, and 7) our approach has still a difficulty in preserving the original sharpness of the texture visible in the original from the FaceStyle algorithm. Such a visual smoothing effect is caused by the fact that the network has parametric nature while the output from FaceStyle represents a non-parametric mosaic of patches that represent exact copies of the original style exemplar. As a future work, we plan to investigate more the possibility to train pixel mapping instead of color information which can enable the formation of the final image using an explicit pixel copy-and-paste operation as in patch-based techniques.

Although our approach delivers stable results when the target does not change considerably and enables rough temporal co-

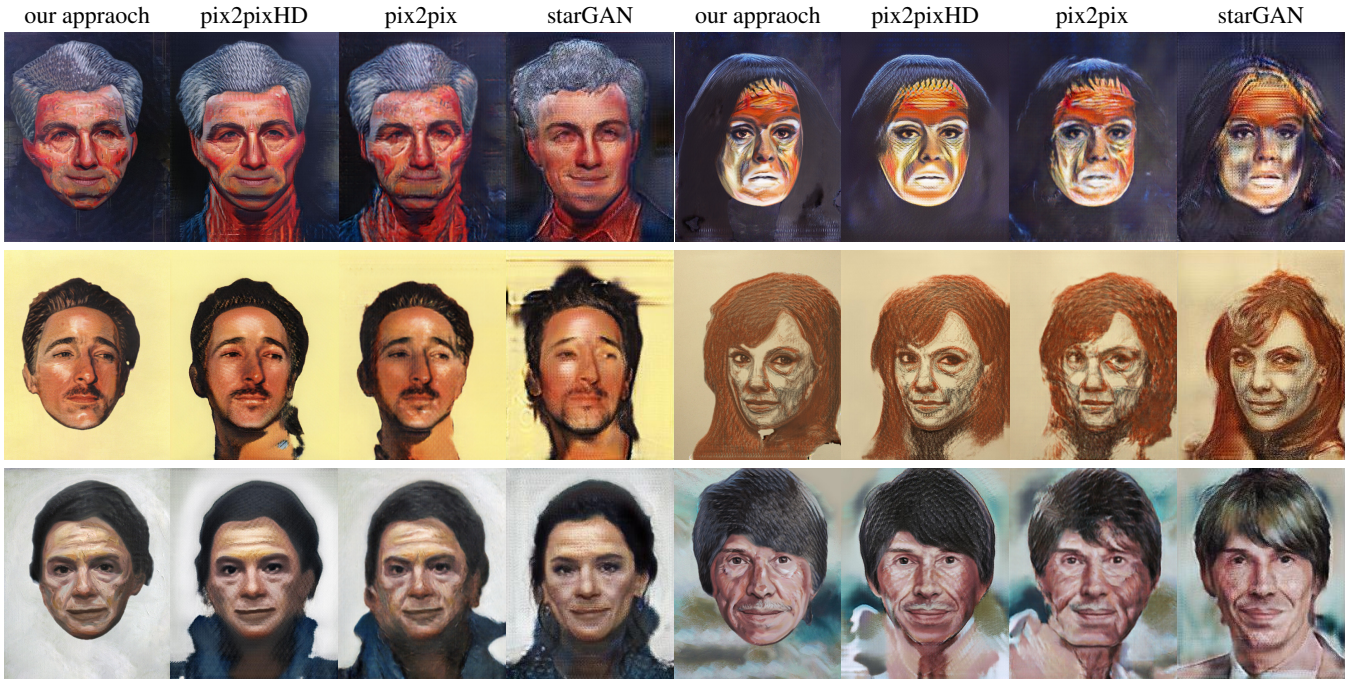

Figure 8: *Comparisons of our approach with current state-of-the-art in image-to-image transation:* pix2pixHD *[WLZ*18a], pix2pix [IZZE17], and* starGAN *[CCK*18]. Note, how our combination of losses and a specific network architecture better preserve the original style exemplar. The corresponding style exemplars are visible in Figures 1, 2, 4, and 9.*

Figure 9: *Comparisons of our approach with current state-of-the-art face stylization methods. Note how our technique can deliver comparable visual quality to the original FaceStyle algorithm of Fišer et al. [FJS*17] while significantly outperforms other concurrent neural-based techniques (Liao et al. [LYY*17], Selim et al. [SED16], and Gatys et al. [GEB16]). Style exemplar:* © Graciela Bombalova-Bogra.

herency for video sequences it still suffers from subtle temporal flicker which can be disturbing in some applications. To gain control over the temporal dynamics an addition of specific temporal smoothness terms similar to those used in video-to-video transfer approaches [WLZ*18b] need to be considered.

6. Conclusion

We present a novel approach for example-based stylization of facial images. Our key idea is to combine a state-of-the-art patch-based synthesis algorithm with a new variant of conditional generative adversarial network. Such a fusion allows us to reach an equilibrium that retains or even improves the visual quality of results produced by the original patch-based approach while increasing its robustness. We compared our combined technique with current state-of-the-art in example-based image stylization as well as in learning-based image-to-image translation methods and reported a considerable quality improvement in both domains. Thanks to the ability to upload our trained generative network into a consumer graphics card we can present the first real-time by-example stylization engine that reaches the visual quality of state-of-the-art techniques tailored to offline processing.

Acknowledgements

We would like to thank Ondřej Jamriška and Jan Keller for help with preparation of training dataset, Šárka Sochorová and Ondřej Texler for recording live demo sessions, and all anonymous reviewers for their insightgful comments. This research began as an internship by David Futschik at Snap. It was funded by Snap and has been supported by the Technology Agency of the Czech Republic under research program TE01020415 (V3C – Visual Computing Competence Center), by the Grant Agency of the Czech Technical University in Prague, grant No. SGS19/179/OHK3/3T/13 (Research of Modern Computer Graphics Methods), and by the Research Center for Informatics, grant No. CZ.02.1.01/0.0/0.0/16_019/0000765.

References

[BPD18] BORA A., PRICE E., DIMAKIS A. G.: AmbientGAN: Generative models from lossy measurements. In *Proceedings of International Conference on Learning Representations* (2018). 2

[BSM*13] BERGER I., SHAMIR A., MAHLER M., CARTER E. J., HODGINS J. K.: Style and abstraction in portrait sketching. *ACM Transactions on Graphics 32*, 4 (2013), 55. 2

[CCK*18] CHOI Y., CHOI M.-J., KIM M., HA J.-W., KIM S., CHOO J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8789–8797. 6, 8

[CLR*04] CHEN H., LIU Z., ROSE C., XU Y., SHUM H.-Y., SALESIN D.: Example-based composite sketching of human portraits. In *Proceedings of International Symposium on Non-photorealistic Animation and Rendering* (2004), pp. 95–102. 2

[CLX*02] CHEN H., LIANG L., XU Y.-Q., SHUM H.-Y., ZHENG N.-N.: Example-based automatic portraiture. In *Proceedings of Asian Conference on Computer Vision* (2002), pp. 171–178. 2

[CLY*17] CHEN D., LIAO J., YUAN L., YU N., HUA G.: Coherent online video style transfer. In *Proceedings of IEEE International Conference on Computer Vision* (2017), pp. 1114–1123. 5

[CLY18] CAO K., LIAO J., YUAN L.: Carigans: Unpaired photo-to-caricature translation. *ACM Transactions on Graphics 37*, 6 (2018), 244:1–244:14. 2

[CXK17] CHEN Q., XU J., KOLTUN V.: Fast image processing with fully-convolutional networks. In *Proceedings of IEEE International Conference on Computer Vision* (2017), pp. 2516–2525. 2

[CZL*02] CHEN H., ZHENG N., LIANG L., LI Y., XU Y.-Q., SHUM H.-Y.: PicToon: A personalized image-based cartoon system. In *Proceedings of ACM International Conference on Multimedia* (2002), pp. 171–178. 2

[DiP07] DIPAOLA S.: Painterly rendered portraits from photographs using a knowledge-based approach. In *Proceedings of SPIE Human Vision and Electronic Imaging* (2007), vol. 6492, pp. 33–43. 2

[FJS*17] FIŠER J., JAMRIŠKA O., SIMONS D., SHECHTMAN E., LU J., ASENTE P., LUKÁČ M., SÝKORA D.: Example-based synthesis of stylized facial animations. *ACM Transactions on Graphics 36*, 4 (2017), 155. 1, 2, 3, 4, 5, 6, 7, 8

[GCLY18] GU S., CHEN C., LIAO J., YUAN L.: Arbitrary style transfer with deep feature reshuffle. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8222–8231. 2

[GEB16] GATYS L. A., ECKER A. S., BETHGE M.: Image style transfer using convolutional neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 2414–2423. 1, 2, 6, 8

[GPAM*14] GOODFELLOW I. J., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A. C., BENGIO Y.: Generative adversarial nets. In *Advances in Neural Information Processing Systems* (2014), pp. 2672–2680. 1, 3

[GRG04] GOOCH B., REINHARD E., GOOCH A.: Human facial illustrations: Creation and psychophysical evaluation. *ACM Transactions on Graphics 23*, 1 (2004), 27–44. 2

[HB17] HUANG X., BELONGIE S. J.: Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of IEEE International Conference on Computer Vision* (2017), pp. 1510–1519. 2

[HJO*01] HERTZMANN A., JACOBS C. E., OLIVER N., CURLESS B., SALESIN D. H.: Image analogies. In *SIGGRAPH Conference Proceedings* (2001), pp. 327–340. 1

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778. 3

[IZZE17] ISOLA P., ZHU J.-Y., ZHOU T., EFROS A. A.: Image-to-image translation with conditional adversarial networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 5967–5976. 1, 3, 4, 6, 8

[JAFF16] JOHNSON J., ALAHI A., FEI-FEI L.: Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of European Conference on Computer Vision* (2016), pp. 694–711. 2, 3, 4

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2014). 4

[LFY*17] LI Y., FANG C., YANG J., WANG Z., LU X., YANG M.-H.: Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems* (2017), pp. 385–395. 2

[LLN11] LI H., LIU G., NGAN K. N.: Guided face cartoon synthesis. *IEEE Transactions on Multimedia 13*, 6 (2011), 1230–1239. 2

[LMH*18] LEHTINEN J., MUNKBERG J., HASSELGREN J., LAINE S., KARRAS T., AITTALA M., AILA T.: Noise2Noise: Learning image restoration without clean data. In *Proceedings of International Conference on Machine Learning* (2018), pp. 2965–2974. 2, 4

[LW16] LI C., WAND M.: Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 2479–2486. 2

[LXZ*18] LU M., XU F., ZHAO H., YAO A., CHEN Y., ZHANG L.: Exemplar-based portrait style transfer. *IEEE Access 6* (2018), 58532–58542. 2

[LYY*17] LIAO J., YAO Y., YUAN L., HUA G., KANG S. B.: Visual attribute transfer through deep image analogy. *ACM Transactions on Graphics 36*, 4 (2017), 120. 1, 2, 6, 8

[LZY*17] LU M., ZHAO H., YAO A., XU F., CHEN Y., ZHANG L.: Decoder network over lightweight reconstructed feature for fast semantic style transfer. In *Proceedings of IEEE International Conference on Computer Vision* (2017), pp. 2488–2496. 2

[MLX*17] MAO X., LI Q., XIE H., LAU R. Y. K., WANG Z., SMOLLEY S. P.: Least squares generative adversarial networks. In *Proceedings of IEEE International Conference on Computer Vision* (2017), pp. 2813–2821. 3

[MZZ10] MENG M., ZHAO M., ZHU S. C.: Artistic paper-cut of human portraits. In *Proceedings of ACM Multimedia* (2010), pp. 931–934. 2

[RDB18] RUDER M., DOSOVITSKIY A., BROX T.: Artistic style transfer for videos and spherical images. *International Journal of Computer Vision 126*, 11 (2018), 1199–1219. 5

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention* (2015), pp. 234–241. 3

[SED16] SELIM A., ELGHARIB M., DOYLE L.: Painting style transfer for head portraits using convolutional neural networks. *ACM Transactions on Graphics 35*, 4 (2016), 129. 1, 2, 6, 8

[SHJ*16] SHEN X., HERTZMANN A., JIA J., PARIS S., PRICE B. L., SHECHTMAN E., SACHS I.: Automatic portrait segmentation for image stylization. *Computer Graphics Forum 35*, 2 (2016), 93–102. 4

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2014). 2, 3

[TL05] TRESSET P., LEYMARIE F. F.: Generative portrait sketching. In *Proceedings of International Conference on Virtual Systems and Multimedia* (2005), pp. 739–748. 2

[UVL16] ULYANOV D., VEDALDI A., LEMPITSKY V. S.: Instance normalization: The missing ingredient for fast stylization. *CoRR abs/1607.08022* (2016). 4

[WCHG13] WANG T., COLLOMOSSE J. P., HUNTER A., GREIG D.: Learnable stroke models for example-based portrait painting. In *Proceedings of British Machine Vision Conference* (2013). 2

[WLZ*18a] WANG T.-C., LIU M.-Y., ZHU J.-Y., TAO A., KAUTZ J., CATANZARO B.: High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8798–8807. 1, 3, 6, 8

[WLZ*18b] WANG T.-C., LIU M.-Y., ZHU J.-Y., YAKOVENKO N., TAO A., KAUTZ J., CATANZARO B.: Video-to-video synthesis. In *Advances in Neural Information Processing Systems* (2018), pp. 1152–1164. 8

[WT09] WANG X., TANG X.: Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 11 (2009), 1955–1967. 2

[WTG*13] WANG N., TAO D., GAO X., LI X., LI J.: Transductive face sketch-photo synthesis. *IEEE Transactions on Neural Networks and Learning Systems 24*, 9 (2013), 1364–1376. 2

[WTG*14] WANG N., TAO D., GAO X., LI X., LI J.: A comprehensive survey to face hallucination. *International Journal of Computer Vision 106*, 1 (2014), 9–30. 2

[WXWT18] WANG C., XU C., WANG C., TAO D.: Perceptual adversarial networks for image-to-image transformation. *IEEE Transactions on Image Processing 27*, 8 (2018), 4066–4079. 1

[XRY*15] XU L., REN J. S. J., YAN Q., LIAO R., JIA J.: Deep edge-aware filters. In *JMLR Workshop and Conference Proceedings* (2015), pp. 1669–1678. 2

[YLL*10] YANG M., LIN S., LUO P., LIN L., CHAO H.: Semantics-driven portrait cartoon stylization. In *Proceedings of International Conference on Image Processing* (2010), pp. 1805–1808. 2

[ZDD*14] ZHANG Y., DONG W., DEUSSEN O., HUANG F., LI K., HU B.-G.: Data-driven face cartoon stylization. In *SIGGRAPH Asia Technical Briefs* (2014), p. 14. 2

[ZKW12] ZHOU H., KUANG Z., WONG K.-Y. K.: Markov weight fields for face sketch synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 1091–1097. 2

[ZPIE17] ZHU J.-Y., PARK T., ISOLA P., EFROS A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of IEEE International Conference on Computer Vision* (2017), pp. 2242–2251. 2

[ZZ11] ZHAO M., ZHU S.-C.: Portrait painting using active templates. In *Proceedings of International Symposium on Non-Photorealistic Animation and Rendering* (2011), pp. 117–124. 2

[ZZP*17] ZHU J.-Y., ZHANG R., PATHAK D., DARRELL T., EFROS A. A., WANG O., SHECHTMAN E.: Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems* (2017), pp. 465–476. 2