FaceBlit: Instant Real-time Example-based Style Transfer to Facial Videos

ANETA TEXLER, ONDŘEJ TEXLER, and MICHAL KUČERA, Czech Technical University in Prague, Faculty of Electrical Engineering, Czech Republic
MENGLEI CHAI, Snap Inc., USA
DANIEL SÝKORA, Czech Technical University in Prague, Faculty of Electrical Engineering, Czech Republic

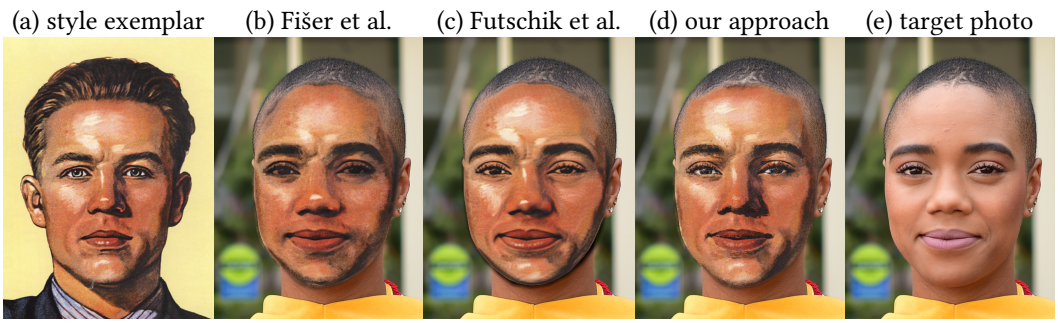

Fig. 1. FaceBlit vs. state-of-the-art—the task at hand is to transfer a style from an exemplar image (a) to a face in the target image (e) while preserving important visual characteristics of the used artistic media in (a) and the identity of the subject in (e). In contrast to current state-of-the-art Fišer et al. [2017] (b) and Futschik et al. [2019] (c), our approach (d) is able to deliver comparable stylization quality and identity preservation without the need to perform costly computation during the synthesis (tens of seconds for Fišer et al.) or lengthy data set generation and training (days for Futschik et al.). Thanks to this advantage our approach can perform instant style transfer to facial videos in real-time even on mobile device. Source style (a) Viktor Ivanovich Govorkov, target photo (f) © Wilson Pumpernickel.

We present FaceBlit—a system for real-time example-based face video stylization that retains textural details of the style in a semantically meaningful manner, i.e., strokes used to depict specific features in the style are present at the appropriate locations in the target image. As compared to previous techniques, our system preserves the identity of the target subject and runs in real-time without the need for large datasets nor lengthy training phase. To achieve this, we modify the existing face stylization pipeline of Fišer et al. [2017] so that it can quickly generate a set of guiding channels that handle identity preservation of the target subject while are still compatible with a faster variant of patch-based synthesis algorithm of Sýkora et al. [2019]. Thanks to these improvements we demonstrate a first face stylization pipeline that can instantly transfer artistic style from a single portrait to the target video at interactive rates even on mobile devices.

CCS Concepts: • **Computing methodologies** → **Non-photorealistic rendering**.

Additional Key Words and Phrases: style transfer, example-based, face stylization, real-time

Authors' addresses: Aneta Texler, aneta.texler@gmail.com; Ondřej Texler, ondrej.texler@gmail.com; Michal Kučera, kucerm22@fel.cvut.cz, Czech Technical University in Prague, Faculty of Electrical Engineering, Karlovo náměstí 13, Praha 2, Czech Republic, 121 35; Menglei Chai, mchai@snap.com, Snap Inc., 2772 Donald Douglas Loop N, Santa Monica, CA, USA, 90405; Daniel Sýkora, sykorad@fel.cvut.cz, Czech Technical University in Prague, Faculty of Electrical Engineering, Karlovo náměstí 13, Praha 2, Czech Republic, 121 35.

ACM Reference Format:
Aneta Texler, Ondřej Texler, Michal Kučera, Menglei Chai, and Daniel Sýkora. 2021. FaceBlit: Instant Real-time Example-based Style Transfer to Facial Videos. *Proc. ACM Comput. Graph. Interact. Tech.* 4, 1 (May 2021), 17 pages. https://doi.org/10.1145/3451270

1 INTRODUCTION

Example-based style transfer becomes recently popular thanks to the significant advances in patch-based synthesis [Fišer et al. 2016; Jamriška et al. 2015] and neural techniques [Gatys et al. 2016; Isola et al. 2017]. As the hardware performance of current mobile devices increases rapidly, it is becoming feasible to perform example-based stylization in real-time even on those small devices.

Recent neural-based style transfer methods [Gatys et al. 2016; Kolkin et al. 2019; Kotovenko et al. 2019b; Li et al. 2017] deliver impressive stylization results. Nevertheless, they tend to omit textural details in the style exemplar that are critical to the preservation of the visual characteristics in the artistic media. Those techniques also do not guarantee a semantically meaningful transfer, i.e., the use of specific local stylization decisions made by an artist in the exemplar image (e.g., use of a certain type of strokes around the mouth area).

On the other hand, although style transfer techniques powered by patch-based methods [Fišer et al. 2016, 2017] can preserve the textural richness and deliver high-quality semantically meaningful results, they are computationally expensive due to their optimization nature. This issue is partially addressed by a faster synthesis algorithm of Sýkora et al. [2019] that provides a real-time approximation to the fully-fledged optimization by leveraging the specific structure of the guiding channels used in the context of face stylization [Fišer et al. 2017]. Despite this great improvement, the time needed to compute the appearance guidance still hinders the real-time performance, which is the reason that Sýkora et al. are not able to demonstrate real-time style transfer that preserves the identity of the target subject.

Recently, Futschik et al. [2019] combine patch-based synthesis [Fišer et al. 2017] with the power of image translation network [Isola et al. 2017] to deliver a first system that enables real-time example-based stylization of facial videos that preserves textural details and is semantically meaningful. Nevertheless, a key limitation of their approach is that for each new style they need to perform lengthy pre-calculation to prepare the dataset and then run yet another time-consuming phase to train the network.

In this paper, we present a method that allows for real-time stylization of an arbitrary facial video using a single stylized exemplar instantly without lengthy pre-calculation. To achieve this, we modify the existing example-based stylization method of Fišer et al. [2017] to compute guidance that is compatible with the fast synthesis method of Sýkora et al. [2019] yet still enables identity preservation of the target subject. To verify the practical utility of the proposed method we implemented the entire stylization pipeline which runs on a moderate mobile device in real-time, and achieves comparable stylization quality with previous techniques.

2 RELATED WORK

The first attempts to perform non-photorealistic rendering [Kyprianidis et al. 2013], i.e., recreating an input image or a video with a specific artistic style, use hand-crafted algorithmic solutions. Some methods compose the final result using a library of predefined assets, e.g., pen and ink strokes [Praun et al. 2001; Salisbury et al. 1997; Snavely et al. 2006], hatching [Breslav et al. 2007], or brush strokes [Hays and Essa 2004; Litwinowicz 1997; Schmid et al. 2011; Zhao and Zhu 2011]. Others try to mimic the given artistic medium by employing physical simulation [Curtis et al. 1997; Haevre et al. 2007; Lu et al. 2012], or using hand-crafted shaders on the GPU [Bénard et al. 2010; Bousseau et al. 2006, 2007; Montesdeoca et al. 2018]. While these techniques are able to produce

faithful stylization to some extent, their use is limited to a certain look given by the predefined visual style.

To address this limitation, example-based techniques start to emerge. Sloan et al. propose the Lit Sphere system [2001] where a hand-drawn illuminated sphere is used to stylize shading on an arbitrary 3D model. To achieve this, Sloan et al. employ a particular variant of texture mapping called environment mapping [Blinn and Newell 1976]. A similar texture mapping-based technique is used to re-project photographs on 3D models by Debevec et al. [1996].

Hertzmann et al. [2001] propose a versatile example-based concept called Image Analogies in which smaller image patches are transferred to the target image from a hand-drawn style exemplar. As compared to texture mapping techniques, this approach can preserve important artistic features such as brush strokes, which are critical to retaining the fidelity of the used artistic media. Another advantage of Image Analogies is their ability to perform semantically meaningful transfer thanks to a set of predefined guiding channels. The power of this general concept is later demonstrated in numerous applications including stylization of fluid simulations [Jamriška et al. 2015], 3D renders [Bénard et al. 2013; Fišer et al. 2016], and facial [Fišer et al. 2017] or arbitrary videos [Jamriška et al. 2019]. However, a crucial drawback of these techniques is that they usually employ costly patch-based synthesis algorithms [Fišer et al. 2016; Kaspar et al. 2015; Wexler et al. 2007]. Even when a GPU is involved, those techniques have difficulties in delivering high-resolution stylized output in real-time. This drawback is critical in scenarios where only a limited computational budget is available, e.g., on a mobile phone.

Sýkora et al. [2019] propose an efficient approximation of patch-based synthesis that bypasses expensive optimization steps achieving real-time stylization even when the computational resources are limited. However, in our face stylization scenario, their method requires a specific type of guidance that is costly to compute and thus the entire stylization cannot run in real-time.

Another successful approximation to patch-based synthesis is recently introduced by Hauptfleisch et al. [2020]. They pre-calculate the latent representation of the stylized image in a sparse set of samples and then merge nearby pre-calculated representations to reconstruct the final stylized image during the interactive session. Although such an approach can deliver similar quality as full-fledged optimization, it requires costly pre-processing and works only on 3D models.

A popular alternative to patch-based style transfer employs neural networks. Gatys et al. [2016] pioneer the idea of back-propagation through the pre-trained VGG network [Simonyan and Zisserman 2014]. They optimize the target image until its VGG responses match the style image as well as the target content. Such optimization is, however, computationally demanding and thus others [Johnson et al. 2016; Ulyanov et al. 2016a,b, 2017; Wang et al. 2017; Wilmot et al. 2017] later propose to pre-calculate a larger dataset in a particular style, and then train a feed-forward network that is able to reproduce the stylized output notably faster. Although those approaches can perform stylization in real-time they still require lengthy pre-processing. Moreover, neural techniques also tend to omit important textural details presented in the original style exemplar and the transfer is not semantically meaningful.

Better performance with respect to preserving textural details and producing semantically meaningful output have image-to-image translation networks [Isola et al. 2017; Zhu et al. 2017a,b]. Those, however, require a large dataset of translation pairs which is usually not accessible in our scenario. Futschik et al. [2019] tried to overcome this limitation by employing results of Fišer et al. [2017] to produce those pairs. Nevertheless, still the drawback is that to add a new style one needs to generate a large dataset and train the network which is computationally costly and time-consuming. Texler et al. [2020b] propose a patch-based training strategy that can notably lower the pre-calculation cost. However, their approach works only on subjects for which training exemplars are provided.

To achieve an arbitrary style transfer using a network trained on unpaired examples, encoder-decoder schemes are proposed [Huang and Belongie 2017; Li et al. 2017; Lu et al. 2017]. In this setup, an encoder, usually a subset of convolutional layers of the VGG network, is used to extract feature representation from both style and content image. These features are then combined and fed through the decoder, which is pre-trained to convert features into the image space. In a similar spirit, more complex encoder-decoder schemes are proposed by Kotovenko et al. [2019a; 2019b]. They are able to convincingly transfer even finer textural details. Nevertheless, as they measure only statistical correlations between the stylized image and the original image, semantically meaningful transfer is not guaranteed.

There are also various successful attempts to combine patch-based synthesis and neural style transfer. Li et al. [2016] search for neural patches in a style image while following the structure of a content image. Liao et al. [2017] extend the original Image Analogies [Hertzmann et al. 2001] framework into the neural domain, where they perform patch-based optimization on feature responses of the VGG network. Their method faithfully reproduces textural details of the given style exemplar, however, it is computationally expensive. Texler et al. [2020a] propose to use patch-based synthesis method on top of the neural-based style transfer approach. In this setting, they are able to generate high-resolution stylized imagery which would be difficult for the original neural network. Their method is able to convincingly preserve important texture details of the style exemplar. But semantically meaningful results are still not guaranteed as the method relies on the output from the underlying neural network.

Recently, few-shot learning techniques [Liu et al. 2019; Wang et al. 2019] and approaches based on deformation transfer [Siarohin et al. 2019a,b] are proposed to animate target photo in real-time using only a single exemplar. A key limitation of these approaches is that they transfer only coarse deformation characteristics while the identity of the subject in the driving video is often omitted.

3 OUR APPROACH

The input to our method is a style exemplar image S of a human portrait and a target face video sequence T. The assumption is that the face changes its expression, moves but is mostly looking towards the camera, and is not occluded by other objects. The output of our method is a stylized sequence O that retains important artistic features of S while preserving the identity of the target subject. Although such an output can already be produced using, e.g., a method of Fišer et al. [2017] a key drawback here is that their approach is suitable only for offline processing. To achieve real-time performance we need to change the way how guiding channels are computed and also replace the slow patch-based synthesis algorithm of Fišer et al. [2016] with its faster variant proposed by Sýkora et al. [2019].

In Fišer et al. [2017] four guiding channels are used to drive the synthesis. A segmentation guide G_{seg} that delineates important facial features by subdividing the face into a set of regions (hair, eyebrows, nose, lips, oral cavity, eyes, and skin) and a positional guide G_{pos} that encodes spatial correspondences between the source and target face. Those two channels ensure semantically meaningful transfer (i.e., strokes used to depict, e.g., eyes in S are used to stylize eyes in T as well). To preserve the identity of the target subject Fišer et al. employs an appearance guide G_{app} which reduces domain gap between the source and target image by equalizing their appearance using the photographic style transfer method of Shih et al. [2014]. Finally, a temporal guide G_{temp} represented by a motion-compensated version of the previously stylized frame is used to enforce temporal consistency.

Since the computation of guiding channels mentioned above takes tens of seconds on a desktop, their use is not tractable for our real-time scenario. Instead, we reduce those four channels into two essential G_{pos} & G_{app} (see Fig. 2), and change their underlying generation algorithms to reduce

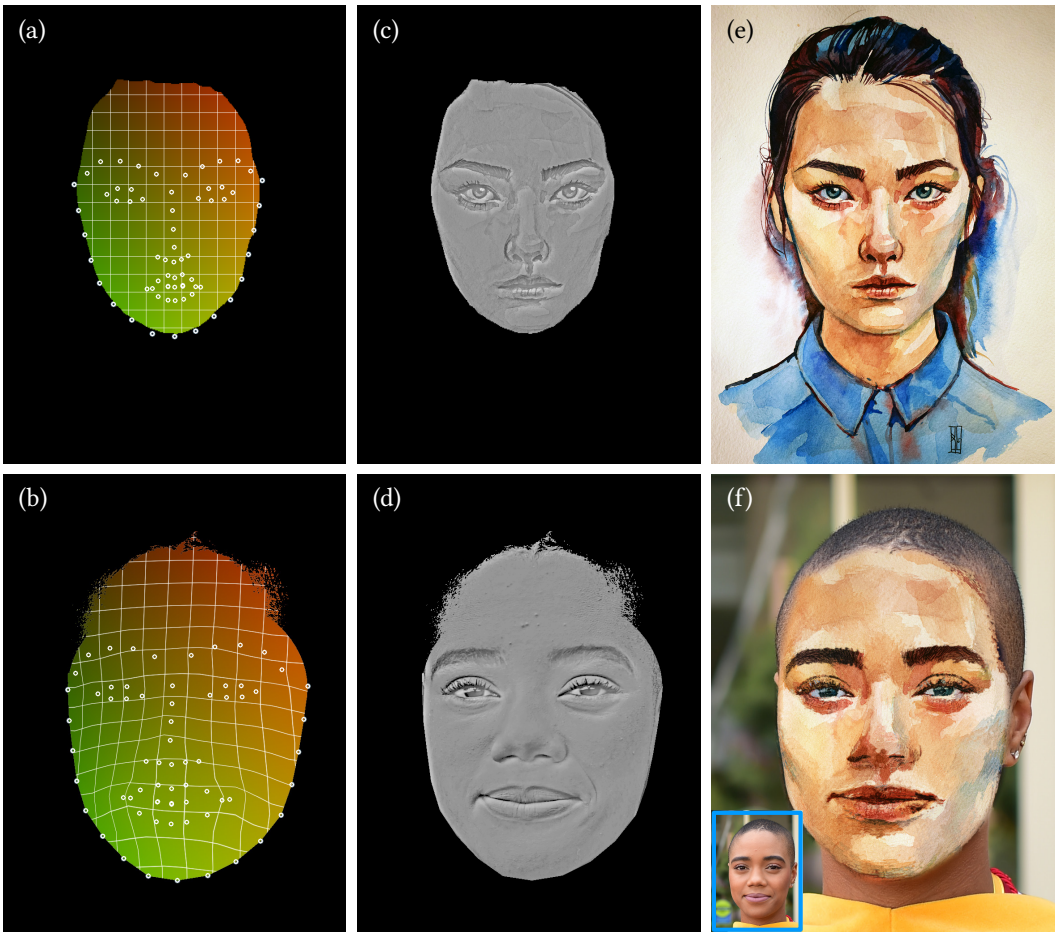

Fig. 2. Overview of the guiding channels used in our technique. The positional guide G_{pos} (a, b) secures the local consistency of the transfer from the style exemplar (e) to the target image (inset in blue). The target positional guide (b) is created by deforming the positional guide of the style image (a) according to the correspondence of facial landmarks, shown as white circles. Note that landmarks and the white grid is shown only for visualization purposes. The appearance guide G_{app} (c, d) encourages the synthesis to preserve subject's identity. See the text and Fig. 4 for detailed explanation of how G_{pos} & G_{app} is computed. Style exemplar (e) © Boris Groh, target photo (f) © Wilson Pumpernickel.

the preparation time to tens of milliseconds. Finally, we demonstrate how to plug those two new guiding channels into a fast synthesis algorithm of Sýkora et al. [2019].

3.1 Positional Guide

A key role of the positional guide G_{pos} is to ensure style consistency, i.e., encourage the synthesis to transfer patches from the source exemplar to a semantically meaningful location in the target image. The existence of the positional guide in the set of guiding channels is also an essential component for the fast synthesis method of Sýkora et al. [2019] which requires one of the guides to provide good localization.

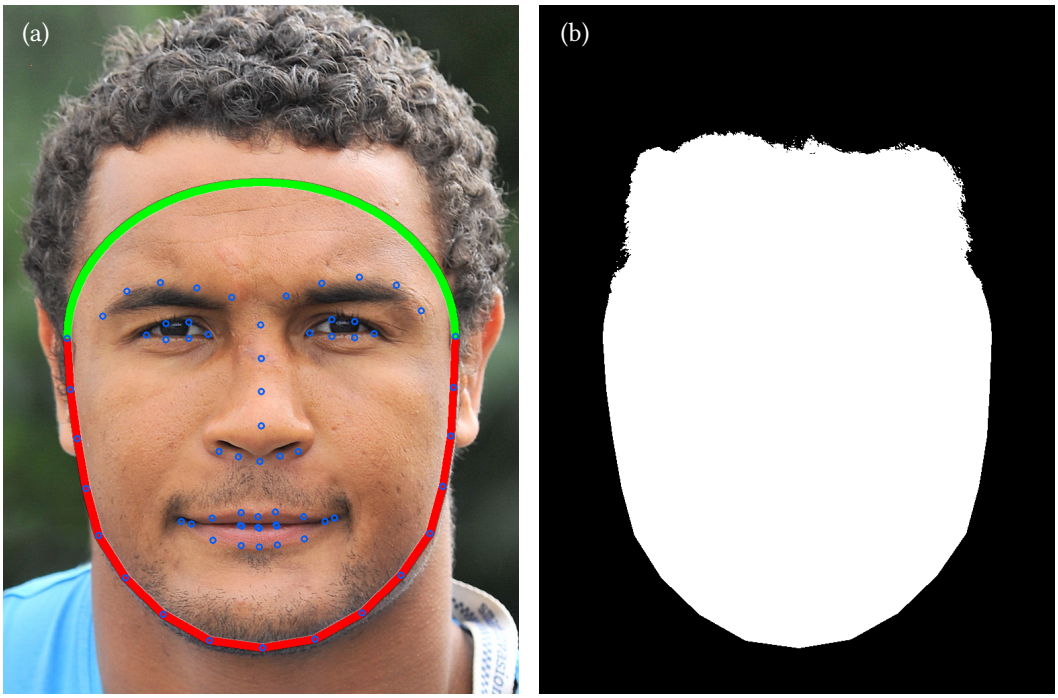

Fig. 3. Given a face (a), we compute a fast approximation of a segmentation mask (b) as follows. We take advantage of detected landmarks visualized as blue circles in (a). We first connect the chin landmarks, red line in (a). Then, we connect left and right uppermost chin landmark using an ellipse, green curve in (b). This gives us the segmentation of a lower and inner face. To include segmentation of forehead, we sample color components along the green curve and use a fast color thresholding operation and connected component analysis to determine the boundary between skin and hair, see the text for details. Target photo (a) © Patrick Subotkiewiez.

Obtaining positional guide G^S_{pos} (Fig. 2a) for the style exemplar S is straightforward. All pixels are simply set to a color determined by their coordinates: x-coordinate corresponds to the red channel, y to the green channel. For $G^{T_i}_{\text{pos}}$ we need to generate an image that encodes a warping field between S and T_i where each target pixel is storing color-coded coordinates of its corresponding pixel in the source image (Fig. 2b). To create $G^{T_i}_{\text{pos}}$ we detect facial landmarks in the style exemplar S as well as in the target frame T_i using the method of Kazemi et al. [2014]. They provide a set of point correspondence from which a warping field between the source and target face can be computed using the moving least-squares method of Schaefer et al. [2006].

Since the style image S is static, facial landmarks can be detected in advance to save computational time. Sometimes landmark detector of Kazemi et al. may fail on artistic images due to the fact that it is trained on real photographs. In such a case, the method of Yaniv et al. [2019] tailored to artistic images could be used instead. In the target frames T_i the detection of landmarks needs to be performed on the fly. Therefore, to increase the detection speed, we subsample the target portrait to half resolution before passing it to the detector. It affects the accuracy negligibly while makes the detection significantly faster.

In contrast to Fišer et al. [2017] we do not explicitly compute G_{seg} to reduce the computational overhead. Instead, we encode a simplified version of the facial mask directly into G_{pos}. In addition

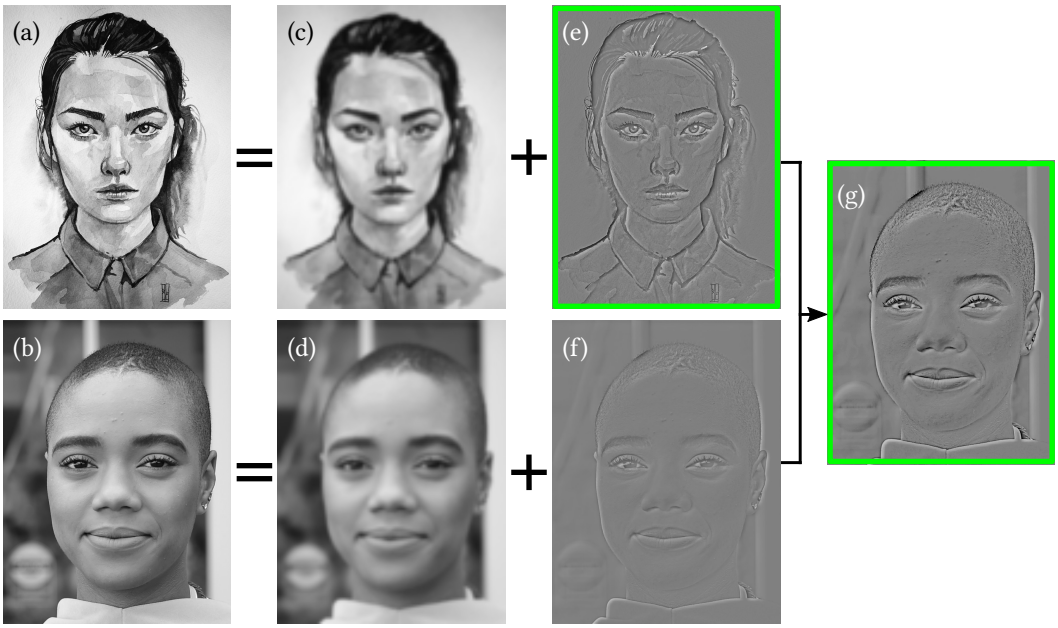

Fig. 4. The process of generating appearance guides G_{app} for the style exemplar S and the target frame T_i. The original images are converted into a grayscale domain (a, b), and filtered using Gaussian blur (c, d). To simulate the result of Laplacian of Gaussian filter (e, f) we subtract the blurred images (c, d) from their originals (a, b). Image (e) is the source part of appearance guide G^S_{app} and to produce its target counterpart $G^{T_i}_{app}$ (g) we modify (f) to match its histogram to that of (e). Style exemplar (a) © Boris Groh, target photo (b) © Wilson Pumpernickel.

to color-coded pixel coordinates, we use the remaining blue channel to store a mask of the facial segment which for the style image S is computed offline using the method of Lee et al. [2020]. For T_i we need a faster algorithm as the target mask is computed on the fly. We use a subset of chin landmarks to define the lower part of the mask boundary. The upper part is constructed by sampling color components of pixels along the upper part of an ellipse going through the left and right uppermost chin landmark. Those samples are then used to perform fast color threshold operation followed by a connected component analysis that extracts the largest region of which upper contour defines the remaining upper boundary of the facial mask (see Fig. 3).

3.2 Appearance Guide

A primary role of appearance guide G_{app} is to preserve the identity of the target subject. In Fišer et al. [2017] a method of Shih et al. [2014] is employed to equalize the target image to have a similar appearance as the source style. However, this approach requires several seconds to compute. The entire Laplacian pyramid for both source and target image needs to be constructed. Then a robust gain mask is computed, applied at each pyramid level. And finally, a pyramid collapsing operation is performed. In our experiments, we found that such a costly operation can be approximated by a computation of only a single pyramid level on which histogram equalization is applied (see Fig. 4).

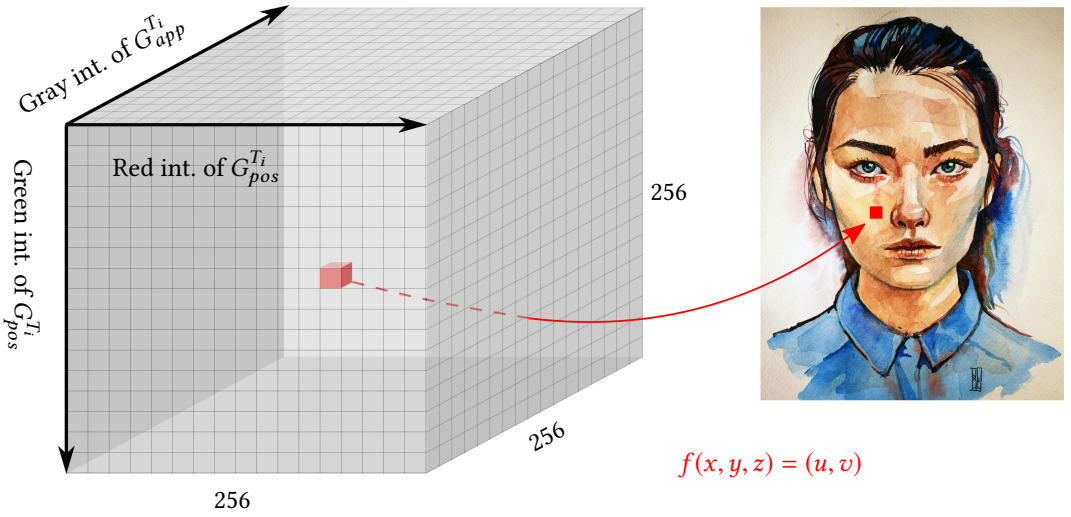

Fig. 5. The utilization of a 3D lookup table to obtain the corresponding source pixel for each target pixel. The cube stores coordinates of the best matching style exemplar pixel for a given red and green channel value in G_{pos} and the gray intensity in G_{app}. It allows to find the corresponding source pixel with complexity $O(1)$ during the synthesis using method of Sýkora et al. [2019]. Style exemplar (right) © Boris Groh.

3.3 Style Transfer

Once the guiding channels for the source image S and the target video frame T_i are computed, the style transfer can be performed using the method of Sýkora et al. [2019]. Since in our case we have more than two values indirectly specifying the corresponding pixel location in the source exemplar (three dimensions for G_{pos} and one dimension for G_{app}), we need to leverage a kind of data structure which for each target pixel q quickly retrieves the closest source pixel p given by the following error metric:

$$E(p, q) = \|G^S_{\text{pos}}(p) - G^{T_i}_{\text{pos}}(q)\|^2 + \lambda |G^S_{\text{app}}(p) - G^{T_i}_{\text{app}}(q)|^2 \quad (1)$$

where λ is weighting the contribution of G_{pos} & G_{app} terms. We first reduce the original 4D mapping into 3D by encoding the blue channel of G_{pos} using zeros in red and green channels (see Fig. 2) and then we pre-calculate a 3D lookup table (see Fig. 5) that will require some additional memory but enables constant retrieval time. Such a 3D lookup table is then plugged into the parallel StyleBlit algorith descrbed in Sýkora et al. [2019] (Algorithm 1).

4 RESULTS

We implemented our approach using Java and C++. For all results presented in the paper we use the following setting of parameters in the StyleBlit algorithm [Sýkora et al. 2019]: $\lambda = 0.2$ and $t = 50$. For each new style exemplar it takes several seconds to pre-calculate necessary data (3D lookup table, landmarks, and guiding channels) before the real-time stylization starts. The most critical is the computation of the 3D lookup table, the structure that stores coordinates pointing to the closest pixels in the source image (see Fig. 5). To obtain the coordinates, entire source image has to be searched. This process is computationally expensive as the search needs to be done for every position of the 3D lookup cube, i.e., 256^3 times. However, we reduce the processing time significantly by restricting the radius for searching the best matching pixel candidate to 20 pixels from the location estimated only by G_{pos}. We empirically verified that for all styles used in our

experiments larger radius do not significantly increase the stylization quality. When a multicore CPU or a GPU is available lookup table pre-calculation can easily be accelerated by subdividing the entire 3D space into a set of smaller cubes that can be evaluated in parallel.

On a half megapixel image our implementation runs at 15 frames per second on *Samsung Galaxy Note8* with CPU *Samsung Exynos 8895, 2.3 GHz*, GPU *Mali G71 MP20* and *6 GB* of RAM. The framerate scales roughly linearly with the increasing number of pixels. On the fly detection of landmarks in the target video frame takes 10 ms, generation of guidance channels 12 ms, style transfer 20 ms, and other miscellaneous steps, (camera handling, frame flipping and rotating, conversions between color spaces, copying data between Java and C++) take 28 ms.

We tested our method with various style exemplars applied on several target faces from FFHQ dataset [Karras et al. 2019] (see Fig. 6) and videos captured on a mobile device (see our supplementary video). Those experiments verified that our method can carry the exemplar's textural details while still being able to respect the target subject's identity. The quality of stylization results is comparable to those produced by the previous offline method of Fišer et al. [2017] as well as real-time method of Futschik et al. [2019] that requires lengthy pre-processing phase (see Fig. 1 and Fig. 11). A detailed banchmark measuring pre-processing and synthesis time for a half megapixel image on 3 GHz Quad-core CPU with Nvidia RTX 2080 GPU is available in Table 1. Note that as compared to Fišer et al. and Futschik et al. our method uses only CPU.

Method	**Pre-calculation**	**Synthesis**
Our approach (CPU only)	10 s	0.05 s
Fišer et al. [2017] (CPU + GPU)	5 s	10 s
Futschik et al. [2019] (CPU + GPU)	2 days	0.06 s

Table 1. Comparison of processing times w.r.t. current state-of-the-art.

In addition to method comparison, we also performed various ablation experiments.

In Fig. 7, we demonstrate the importance of using both the G_{pos} & G_{app} guidance channels. The absence of G_{pos} may cause that coherent chunks from style exemplar are transferred to wrong locations in the target portrait, (see Fig. 7c, d). Without G_{app}, the subject's identity is not preserved well (see, e.g., wrong eyebrows or the absence of wrinkles in Fig. 7e, f). When using both guides, stylized results faithfully represent artistic medium, the transfer is semantically meaningful, and the identity of the target subject is well-preserved (see Fig. 7g, h).

In Fig. 8 we show the necessity of the histogram matching operation during the generating of the target appearance guide G^T_{app}. Without matching the appearance guides' histograms, the error E overcomes the threshold t too soon which leads to notably smaller chunks and the result may seem blurry (see Fig. 8).

We also tried to execute our algorithm with the same G^T_{app} as described in the original approach of Fišer et al. [2017] (see Fig. 9). It is visible that their more sophisticated G^T_{app} preserves the subject's identity a bit better, nevertheless, it is notably slower to compute.

5 EXTENSIONS

A visible limitation of our approach when compared to current state-of-the-art is the absence of hair stylization (c.f. Fig. 11). Although the face parsing network of Lee et al. [2020] can be used to estimate hair mask its computational overhead is too demanding to preserve real-time response. Also the computation of positional guide could be complicated when the shapes of source and target hair segments differ significantly. The resulting warping field may violate good localization

Fig. 6. FaceBlit applied on several target subjects (leftmost column), using various style exemplars (topmost row). Style exemplars: (a) © Boris Groh, (b) Viktor Ivanovich Govorkov, (c) © Matthew Ivan Cherry (HAT, oil on canvas, 48" x 48", 2011), (d, e) © Adrian Morgan, (f) Peter Zelizňák (sculpture by Stanislav Mikuš), target photos: (g) PFA SEAL, (h) © Ajuntament de Sabadell, (i) © Raziel Janeway, (j) lam_anh2005.

property of the positional guide which is crucial for the method of Sýkora et al. [2019] to produce reasonable stylization results.

To alleviate this drawback, we implemented a hybrid method that uses our new face stylization approach to bring an existing portrait painting to life while adapting the identity of the portrayed person to the one seen in the target video (see Fig. 10 and our supplementary video). In this extension we separate the style image into a set of segments (face, hair, beard, torso, and background). These segments are processed independently and then stiched together to form the final output frame. The facial segment is stylized using the algorithm described in this paper. For hair, beard (if applicable), and torso segments, we use moving least-squares deformation [Schaefer et al. 2006] driven by a set of facial landmarks (c.f. Fig. 10).

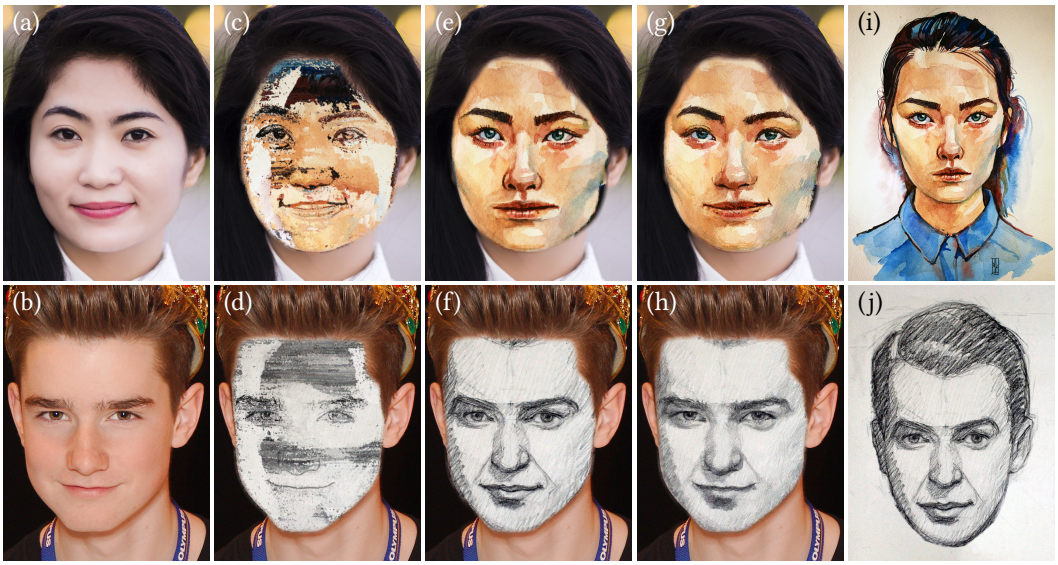

Fig. 7. Importance of individual guidance channels. The positional guide G_{pos} is essential. Its absence (c, d) causes that the chunks from the style are not transferred in a semantically meaningful way. Without the appearance guide G_{app}, the identity of target subjects (a, b) is not preserved well (e, f). The full guidance (g, h) secures the local consistency of style transfer while retaining the target subject's identity. Style exemplars: (i) © Boris Groh, (j) © Adrian Morgan, target photos: (a) © LEMON Studio, (b) © Mark Peers.

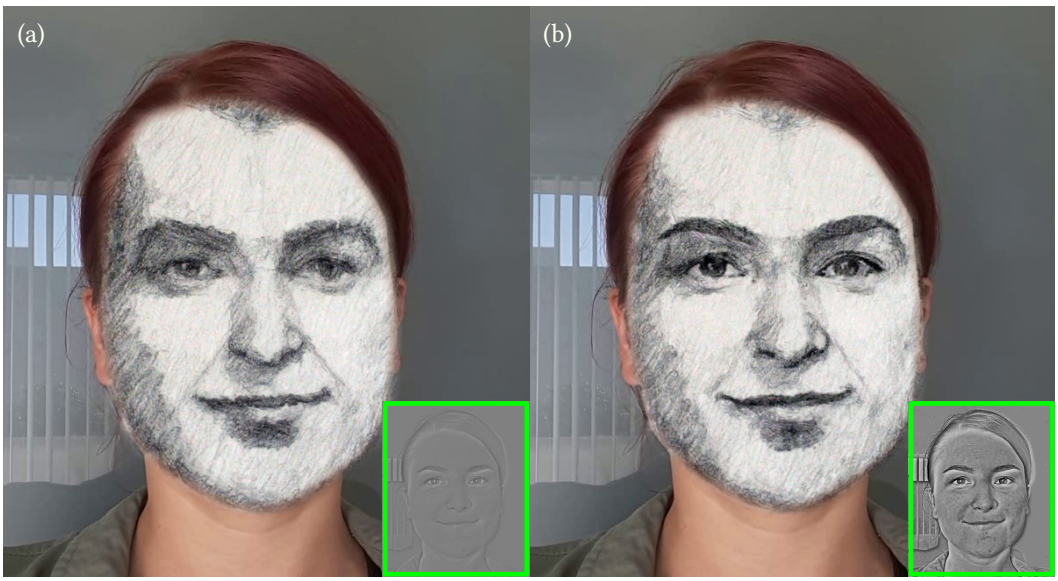

Fig. 8. Importance of the histogram matching phase during the generation of the target appearance guide G^T_{app}. Without the histogram matching, the subject's identity is not preserved well, and the result may seem blurry. See (a) and its respective appearance guide in green inset. After equalizing histograms, the gain in quality is significant. See (b) and the green inset.

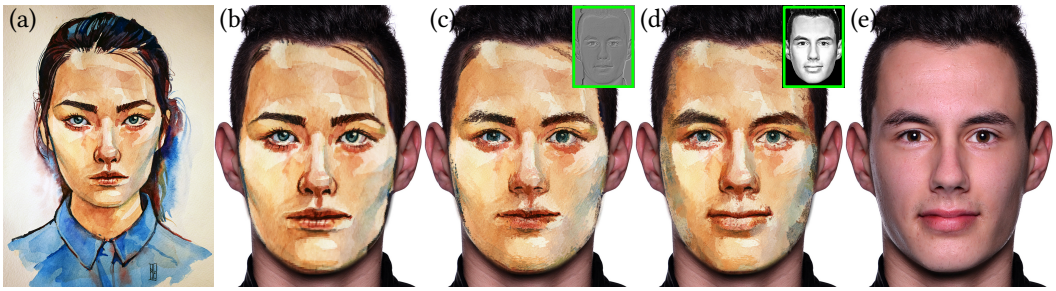

Fig. 9. Comparison of using our appearance guide with the one proposed in Fišer et al. [2017]—style from the exemplar (a) is transferred to the target image (e). A stylization result without appearance guide (b), with appearance guide generated by our method (c), and with appearance guide generated by Fišer et al. (d). Note, how the identity of the target subject is bit less pronounced as compared to the solution of Fišer et al., which however is orders of magnitude slower than ours. Style exemplar (a) © Boris Groh, target photo (e) SKV Florbal.

6 LIMITATIONS AND FUTURE WORK

Although we show that our approach can deliver comparable visual quality with significantly lower computational overhead than the current state-of-the-art, some limitations stem from this performance gain.

A compromise we accepted in our real-time solution is the omission of explicit guidance that allows to control the level of temporal coherence. Although the flickering our approach is producing resembles temporal dynamics of hand-colored animations and can be perceived as an important feature (c.f. Fišer et al. [2014]), some sort of control over its behavior would be valuable since it may become disturbing after a while. To control the strength of temporal flickering Sýkora et al. [2019] propose to lower the threshold t of their fast stylization algorithm which in fact leads to smaller copied exemplar chunks and thus become close to texture mapping scenario which breaks the planarity of brush strokes present in the original style exemplar. This problem opens an interesting direction for future work.

Also, the addition of appearance guide causes the overall guidance to become a bit more discontinuous when compared to the case of clean positional guide which better suits fast stylization method of Sýkora et al. [2019]. Due to this reason the size of transferred chunks can be notably smaller and thus cause suppression of artistic features that have larger scale in the original style exemplar (see, e.g., Fig. 7f vs. 7h).

Lastly, our approach shares similar limitations as other techniques that use guided patch-based synthesis [Fišer et al. 2017; Futschik et al. 2019; Sýkora et al. 2019]. The style exemplar needs to have a compatible scale with the target image otherwise artifacts may appear (see, e.g., Fig. 13 in [Sýkora et al. 2019]). Patch-based synthesis also encounters difficulties when adapting to different lighting conditions or an absence of important features (e.g., wrinkles or moustach) that are present in the target image, however, are missing in the style exemplar or vice versa. A viable avenue for future work could be to alter between a set of exemplars drawn in a similar style that would better suit the target image (e.g., various lighting directions, man/woman, old/young, etc.).

7 CONCLUSION

We present the first algorithm enabling instant real-time example-based stylization of facial videos with semantically meaningful output. It allows retaining the notion of original artistic media while preserving the target subject's identity without the need to perform lengthy pre-calculation or

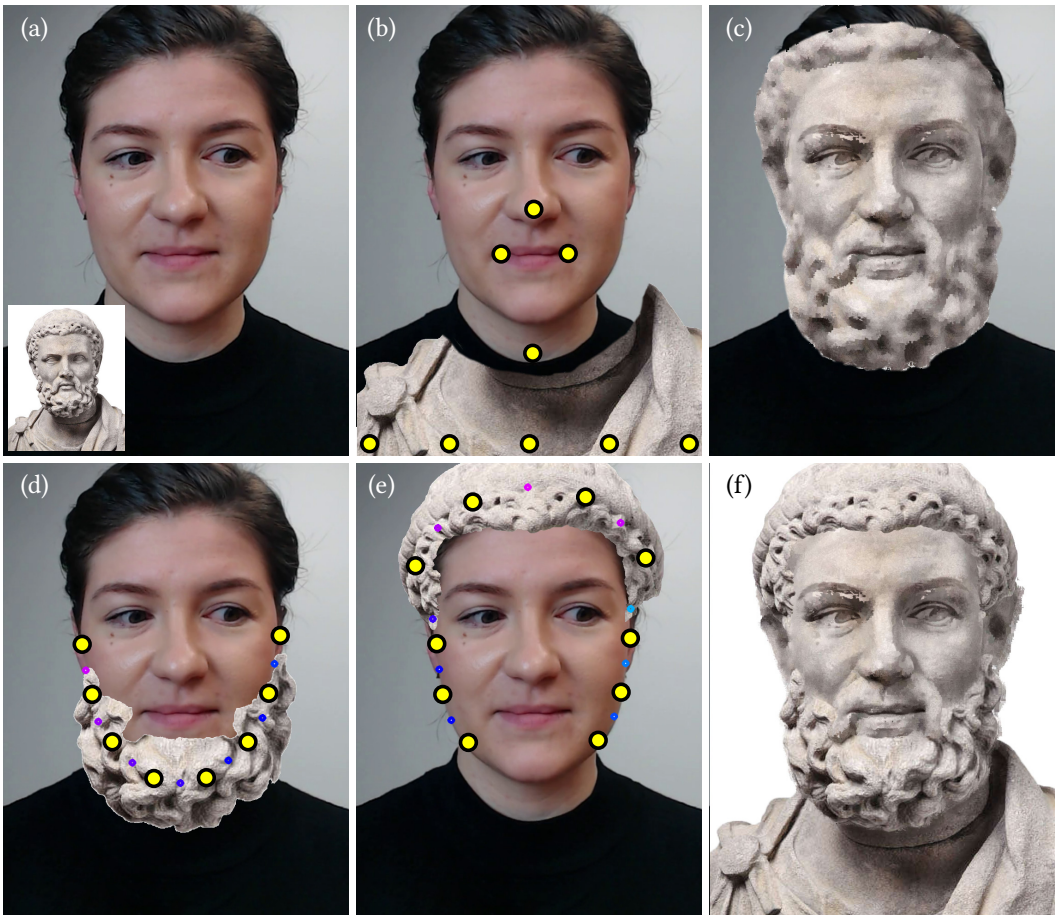

Fig. 10. An example of a hybrid approach where the aim is to stylize a person in the video (a) to look like the statue in the inset reassembling her identity. To do that we subdivide the statue into a set of separate layers: torso (b), face (c), beard (d), and hair (e). The facial layer (c) is animated using our approach while for the torso (b), beard (d), and hair (e) layer we use moving least-squares deformation [Schaefer et al. 2006] driven by a set of control points (yellow dots) of which position is derived from detected landmarks. Such a set of deformed and stylized layers is then blended in a predefined depth order to produce the final composition (f). See our supplementary video for this example in montion. Style exemplar (a) © Country French Interiors, target photo (a) Šárka Sochorová.

training. We have shown how to quickly calculate a basic set of guiding channels and plug them into a fast variant of patch-based synthesis algorithm to deliver interactive style transfer even on mobile devices. Despite the fact, our approach reduces the computational overhead significantly in contrast to the current state-of-the-art, it still provides a comparable stylization quality and identity preservation. We leverage this advantage in an interactive application running on a mobile device where we instantly animate an existing hand-drawn portrait mimicking the identity of the target subject.

Fig. 11. Comparison of our method with current state-of-the-art: style from an exemplar (a, c, e) is transferred to the target photo (b, d, f) using the method of Fišer et al. [2017] (g), Futschik et al. [2019] (h), and our approach (i). Note, how our approach produces comparable stylization quality while is notably faster than the method of Fišer et al. and does not require lengthy pre-calculation contrary to Futschik et al. A limitation of our method is that it does not support hair stylization. Style exemplars: (a) © Matthew Ivan Cherry (HAT, oil on canvas, 48" x 48", 2011), (c, e) © Adrian Morgan, target photos: (b) © MPCA Photos, (d) © LEMON Studio, (f) © Patrick Subotkiewiez.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable feedback and comments. This research was supported by Snap, 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

Pierre Bénard, Forrester Cole, Michael Kass, Igor Mordatch, James Hegarty, Martin Sebastian Senn, Kurt Fleischer, Davide Pesare, and Katherine Breeden. 2013. Stylizing Animation By Example. *ACM Transactions on Graphics* 32, 4 (2013), 119.

Pierre Bénard, Ares Lagae, Peter Vangorp, Sylvain Lefebvre, George Drettakis, and Joëlle Thollot. 2010. A Dynamic Noise Primitive for Coherent Stylization. *Computer Graphics Forum* 29, 4 (2010), 1497–1506.

James F. Blinn and Martin E. Newell. 1976. Texture and Reflection in Computer Generated Images. *Commun. ACM* 19, 10 (1976), 542–547.

Adrien Bousseau, Matthew Kaplan, Joëlle Thollot, and François X. Sillion. 2006. Interactive watercolor rendering with temporal coherence and abstraction. In *Proceedings of International Symposium on Non-Photorealistic Animation and Rendering*. 141–149.

Adrien Bousseau, Fabrice Neyret, Joëlle Thollot, and David Salesin. 2007. Video watercolorization using bidirectional texture advection. *ACM Transactions on Graphics* 26, 3 (2007), 104.

Simon Breslav, Karol Szerszen, Lee Markosian, Pascal Barla, and Joëlle Thollot. 2007. Dynamic 2D patterns for shading 3D scenes. *ACM Transactions on Graphics* 26, 3 (2007), 20.

Cassidy J. Curtis, Sean E. Anderson, Joshua E. Seims, Kurt W. Fleischer, and David H. Salesin. 1997. Computer-generated watercolor. In *SIGGRAPH Conference Proceedings*. 421–430.

Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. 1996. Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image-Based Approach. In *SIGGRAPH Conference Proceedings*. 11–20.

Jakub Fišer, Ondřej Jamriška, Michal Lukáč, Eli Shechtman, Paul Asente, Jingwan Lu, and Daniel Sýkora. 2016. StyLit: Illumination-Guided Example-Based Stylization of 3D Renderings. *ACM Transactions on Graphics* 35, 4 (2016), 92.

Jakub Fišer, Ondřej Jamriška, David Simons, Eli Shechtman, Jingwan Lu, Paul Asente, Michal Lukáč, and Daniel Sýkora. 2017. Example-Based Synthesis of Stylized Facial Animations. *ACM Transactions on Graphics* 36, 4 (2017), 155.

Jakub Fišer, Michal Lukáč, Ondřej Jamriška, Martin Čadík, Yotam Gingold, Paul Asente, and Daniel Sýkora. 2014. Color Me Noisy: Example-based Rendering of Hand-colored Animations with Temporal Noise Control. *Computer Graphics Forum* 33, 4 (2014), 1–10.

David Futschik, Menglei Chai, Chen Cao, Chongyang Ma, Aleksei Stoliar, Sergey Korolev, Sergey Tulyakov, Michal Kučera, and Daniel Sýkora. 2019. Real-Time Patch-Based Stylization of Portraits Using Generative Adversarial Network. In *Proceedings of the ACM/EG Expressive Symposium*. 33–42.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 2414–2423.

William Van Haevre, Tom Van Laerhoven, Fabian Di Fiore, and Frank Van Reeth. 2007. From Dust Till Drawn: A real-time bidirectional pastel simulation. *The Visual Computer* 23, 9–11 (2007), 925–934.

Filip Hauptfleisch, Ondřej Texler, Aneta Texler, Jaroslav Křivánek, and Daniel Sýkora. 2020. StyleProp: Real-time Example-based Stylization of 3D Models. *Computer Graphics Forum* 39, 7 (2020), 575–586.

James Hays and Irfan A. Essa. 2004. Image and Video Based Painterly Animation. In *Proceedings of International Symposium on Non-Photorealistic Animation and Rendering*. 113–120.

Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. 2001. Image Analogies. In *SIGGRAPH Conference Proceedings*. 327–340.

Xun Huang and Serge J. Belongie. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. *Proceedings of IEEE International Conference on Computer Vision* (2017), 1510–1519.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 5967–5976.

Ondřej Jamriška, Jakub Fišer, Paul Asente, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. 2015. LazyFluids: Appearance Transfer for Fluid Animations. *ACM Transactions on Graphics* 34, 4 (2015), 92.

Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. 2019. Stylizing Video by Example. *ACM Transactions on Graphics* 38, 4 (2019), 107.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Proceedings of European Conference on Computer Vision*. 694–711.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 4401–4410.

Alexandre Kaspar, Boris Neubert, Dani Lischinski, Mark Pauly, and Johannes Kopf. 2015. Self Tuning Texture Optimization. *Computer Graphics Forum* 34, 2 (2015), 349–360.

Vahid Kazemi and Josephine Sullivan. 2014. One Millisecond Face Alignment with an Ensemble of Regression Trees. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 1867–1874.

Nicholas I. Kolkin, Jason Salavon, and Gregory Shakhnarovich. 2019. Style Transfer by Relaxed Optimal Transport and Self-Similarity. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 10051–10060.

Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Björn Ommer. 2019a. Content and Style Disentanglement for Artistic Style Transfer. In *Proceedings of IEEE International Conference on Computer Vision*. 4421–4430.

Dmytro Kotovenko, Artsiom Sanakoyeu, Pingchuan Ma, Sabine Lang, and Björn Ommer. 2019b. A Content Transformation Block for Image Style Transfer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 10032–10041.

Jan Eric Kyprianidis, John Collomosse, Tinghuai Wang, and Tobias Isenberg. 2013. State of the "Art": A Taxonomy of Artistic Stylization Techniques for Images and Video. *IEEE Transactions on Visualization and Computer Graphics* 19, 5 (2013), 866–885.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 5548–5557.

Chuan Li and Michael Wand. 2016. Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 2479–2486.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal Style Transfer via Feature Transforms. In *Advances in Neural Information Processing Systems*. 385–395.

Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual Attribute Transfer Through Deep Image Analogy. *ACM Transactions on Graphics* 36, 4 (2017), 120.

Peter Litwinowicz. 1997. Processing Images and Video for an Impressionist Effect. In *SIGGRAPH*. 407–414.

Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. 2019. Few-Shot Unsupervised Image-to-Image Translation. In *Proceedings of IEEE International Conference on Computer Vision*. 10551–10560.

Cewu Lu, Li Xu, and Jiaya Jia. 2012. Combining sketch and tone for pencil drawing production. In *Proceedings of International Symposium on Non-Photorealistic Animation and Rendering*. 65–73.

Ming Lu, Hao Zhao, Anbang Yao, Feng Xu, Yurong Chen, and Xiang Lin. 2017. Decoder Network over Lightweight Reconstructed Feature for Fast Semantic Style Transfer. *Proceedings of IEEE International Conference on Computer Vision* (2017), 2488–2496.

Santiago E Montesdeoca, Hock Soon Seah, Amir Semmo, Pierre Bénard, Romain Vergne, Joëlle Thollot, and Davide Benvenuti. 2018. MNPR: A Framework for Real-Time Expressive Non-Photorealistic Rendering of 3D Computer Graphics. In *Proceedings of The Joint Symposium on Computational Aesthetics and Sketch Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering*. 11.

Emil Praun, Hugues Hoppe, Matthew Webb, and Adam Finkelstein. 2001. Real-Time Hatching. In *SIGGRAPH*. 581–586.

Michael P. Salisbury, Michael T. Wong, John F. Hughes, and David H. Salesin. 1997. Orientable Textures for Image-based Pen-and-ink Illustration. In *SIGGRAPH Conference Proceedings*. 401–406.

Scott Schaefer, Travis McPhail, and Joe Warren. 2006. Image Deformation Using Moving Least Squares. *ACM Transactions on Graphics* 25, 3 (2006), 533–540.

Johannes Schmid, Martin Sebastian Senn, Markus Gross, and Robert W. Sumner. 2011. OverCoat: an implicit canvas for 3D painting. *ACM Transactions on Graphics* 30, 4 (2011), 28.

Yi-Chang Shih, Sylvain Paris, Connelly Barnes, William T. Freeman, and Frédo Durand. 2014. Style Transfer for Headshot Portraits. *ACM Transactions on Graphics* 33, 4 (2014), 148.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019a. Animating Arbitrary Objects via Deep Motion Transfer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 2377–2386.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019b. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems*. 7135–7145.

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

Peter-Pike J. Sloan, William Martin, Amy Gooch, and Bruce Gooch. 2001. The Lit Sphere: A Model for Capturing NPR Shading from Art. In *Proceedings of Graphics Interface*. 143–150.

Noah Snavely, C. Lawrence Zitnick, Sing Bing Kang, and Michael F. Cohen. 2006. Stylizing 2.5-D video. In *Proceedings of International Symposium on Non-Photorealistic Animation and Rendering*. 63–69.

Daniel Sýkora, Ondřej Jamriška, Ondřej Texler, Jakub Fišer, Michal Lukáč, Jingwan Lu, and Eli Shechtman. 2019. StyleBlit: Fast Example-Based Stylization with Local Guidance. *Computer Graphics Forum* 38, 2 (2019), 83–91.

Ondřej Texler, David Futschik, Jakub Fišer, Michal Lukáč, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. 2020a. Arbitrary Style Transfer Using Neurally-Guided Patch-Based Synthesis. *Computers & Graphics* 87 (2020), 62–71.

Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Š. Sochorová, Menglei Chai, Sergey Tulyakov, and Daniel Sýkora. 2020b. Interactive Video Stylization Using Few-Shot Patch-Based Training. *ACM Transactions on Graphics* 39, 4 (2020), 73.

Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S. Lempitsky. 2016a. Texture Networks: Feed-Forward Synthesis of Textures and Stylized Images. In *ICML*, Vol. 48. 1349–1357.

Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. 2016b. Instance Normalization: The Missing Ingredient for Fast Stylization. *CoRR* abs/1607.08022 (2016).

Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. 2017. Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 4105–4113.

Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. 2019. Few-shot Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems*. 5014–5025.

Xin Wang, Geoffrey Oxholm, Da Zhang, and Yuan-Fang Wang. 2017. Multimodal Transfer: A Hierarchical Deep Convolutional Neural Network for Fast Artistic Style Transfer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 7178–7186.

Yonatan Wexler, Eli Shechtman, and Michal Irani. 2007. Space-Time Completion of Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 3 (2007), 463–476.

Pierre Wilmot, Eric Risser, and Connelly Barnes. 2017. Stable and Controllable Neural Texture Synthesis and Style Transfer Using Histogram Losses. *CoRR* abs/1701.08893 (2017).

Jordan Yaniv, Yael Newman, and Ariel Shamir. 2019. The Face of Art: Landmark detection and geometric style in portraits. *ACM Transactions on Graphics* 38, 4 (2019), 60.

Mingtian Zhao and Song-Chun Zhu. 2011. Portrait Painting Using Active Templates. In *Proceedings of International Symposium on Non-Photorealistic Animation and Rendering*. 117–124.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017a. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of IEEE International Conference on Computer Vision*. 2242–2251.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. 2017b. Toward Multimodal Image-to-Image Translation. In *Advances in Neural Information Processing Systems*. 465–476.