Example-Based Synthesis of Stylized Facial Animations

JAKUB FIŠER, Czech Technical University in Prague, Faculty of Electrical Engineering
ONDŘEJ JAMRIŠKA, Czech Technical University in Prague, Faculty of Electrical Engineering
DAVID SIMONS, Adobe Research
ELI SHECHTMAN, Adobe Research
JINGWAN LU, Adobe Research
PAUL ASENTE, Adobe Research
MICHAL LUKÁČ, Adobe Research
DANIEL SÝKORA, Czech Technical University in Prague, Faculty of Electrical Engineering

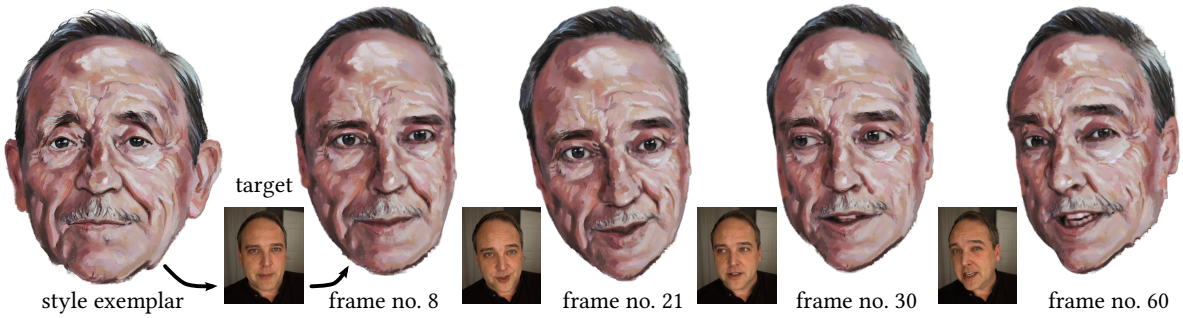

Fig. 1. The painterly style from an exemplar painting (far left) is transferred to a target video sequence (bottom). Exemplar painting: © *Graciela Bombalova-Bogra*. Video sequence: © *Ted Forbes* via YouTube.

We introduce a novel approach to example-based stylization of portrait videos that preserves both the subject's identity and the visual richness of the input style exemplar. Unlike the current state-of-the-art based on neural style transfer [Selim et al. 2016], our method performs non-parametric texture synthesis that retains more of the local textural details of the artistic exemplar and does not suffer from image warping artifacts caused by aligning the style exemplar with the target face. Our method allows the creation of videos with less than full temporal coherence [Ruder et al. 2016]. By introducing a controllable amount of temporal dynamics, it more closely approximates the appearance of real hand-painted animation in which every frame was created independently. We demonstrate the practical utility of the proposed solution on a variety of style exemplars and target videos.

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

Additional Key Words and Phrases: Style Transfer, Hand-Drawn Animation

ACM Reference format:
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 Trans. Graph.* 36, 4, Article 155 (July 2017), 11 pages.
DOI: http://dx.doi.org/10.1145/3072959.3073660

DOI: http://dx.doi.org/10.1145/3072959.3073660

1 INTRODUCTION

Recently, neural-based style transfer has become extremely popular thanks to the seminal work of Gatys et al. [2016] and its numerous publicly available implementations like *DeepArt* and *Prisma*. Selim et al. [2016] extended this technique to provide better results when stylizing head portraits. In their system, additional spatial constraints improve the resemblance between the stylized portrait and its real counterpart. They align the style image to the target photo and compute a set of gain maps to modify the response of the neural network in order to suppress the local differences in appearance.

Although their neural-based style transfer produces impressive results on various styles, it has one key limitation. For styles that contain rich textural information, the method tends to distort local visual features. In some cases, the overall appearance of the synthesized output becomes notably different from the original style exemplar (see Fig. 2, top row). This issue stems from the original method of Gatys et al. being based on a variant of parametric texture synthesis [Portilla and Simoncelli 2000], which is known to produce such artifacts [Efros and Freeman 2001]. Fišer at al. [2016] demonstrated that non-parametric texture synthesis can alleviate this issue, but it is not clear how to apply their analogy-based style transfer technique, designed for 3D rendering, to portrait stylization.

Another issue with Selim et al.'s approach is that it requires perfect alignment (warping) of the source style with the target photo. When the facial proportions of the stylized portrait differ considerably

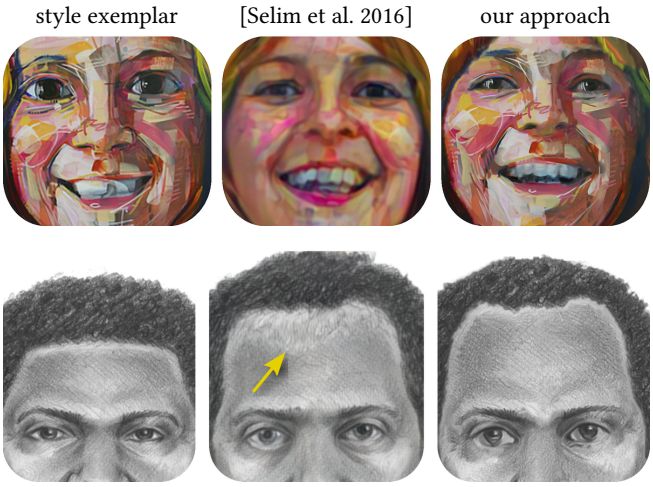

Fig. 2. Artifacts of neural-based style transfer. Top row: Selim et al. [2016] tend to suppress important local visual features (middle), resulting in an overall appearance that differs significantly from the original style exemplar (left). Our approach reproduces them faithfully (right). Bottom row: Selim et al. introduce warping artifacts (elongation and smearing) caused by alignment of the style exemplar (left) to the target face (middle). Our approach can transfer the style without the need to warp the exemplar image (right). Style exemplars (top to bottom): © *Gwenn Seemel*, *NYPD*

from those in the target image, noticeable textural distortion can occur in the stylized output (see Fig. 2, bottom row).

Lastly, similar to other work [Ruder et al. 2016], Selim et al. aim to fully preserve temporal coherence when stylizing a video sequence, resulting in a distinctive look in which style elements appear texture-mapped to the subject's face. This contrasts with the appearance of real hand-painted animations, which exhibit a certain level of temporal flickering – see, for example, the works of Bill Plympton, Aleksandr Petrov's *The Old Man and the Sea*, or the recently produced feature movie *Loving Vincent*[1]. Hand-painted animations tend to preserve temporal coherence only at a coarse level because the physical properties of artistic media make high-frequency details very difficult to control. Creating the perception that the animation was hand-painted frame by frame requires a certain amount of temporal incoherence [Fišer et al. 2014].

We introduce a novel approach to portrait style transfer that preserves both the identity of the target subject and the textural richness of the style exemplar. Our key contribution is an algorithm that, when given a photo-style pair, automatically generates a set of meaningful guiding channels that can be directly used as input to a state-of-the-art non-parametric texture synthesis framework [Fišer et al. 2016]. In contrast to neural-based style transfer [Selim et al. 2016], our technique does not require explicit image warping and preserves low-level textural details that are important for the chosen artistic media. Moreover, in video stylization, our approach introduces a controllable amount of temporal flickering in the spirit of Color Me Noisy [Fišer et al. 2014], which helps to deliver a fully hand-painted look. We demonstrate the practical utility of the proposed style-transfer method on various examples and perform result comparisons to confirm that our technique alleviates drawbacks of previous approaches.

[1]http://lovingvincent.com

2 RELATED WORK

Stylizing photographs and videos is one of the key challenges of non-photorealistic rendering [Kyprianidis et al. 2013]. Besides general purpose approaches [Bénard et al. 2013; DeCarlo and Santella 2002; Hays and Essa 2004; Hertzmann et al. 2001; Winnemöller et al. 2006; Zeng et al. 2009] there are techniques that take into account specific properties of head portraits. They can be divided into filtering-based and example-based techniques.

Filtering-based techniques [DiPaola 2007; Gooch et al. 2004; Tresset and Leymarie 2005; Yang et al. 2010] combine image processing methods like thresholding, segmentation, posterization, edge-detection, saliency measurement, and blurring, using parameter settings tailored to specific features of portrait images. Although they can achieve attractive results, the range of what they can create is limited by the visual properties of the image processing filters being used.

Example-based techniques alleviate the disadvantage of limited visual range by letting the user provide an arbitrary style exemplar. A typical approach is to decompose the face into visually important parts like eyes, nose, mouth, and hair, and to let the artist stylize them separately. The algorithm then spatially distributes and composes those exemplars to meet the specific proportions of the target face [Chen et al. 2002, 2004, 2002; Meng et al. 2010; Zhang et al. 2014]. Although they often provide compelling results their key drawbacks are that the observer often recognises individual reused templates, and without additional exemplars it is hard to achieve a truer representation of specific visual features of the target face. Other example-based techniques use multiscale Markov Random Fields [Li et al. 2011; Wang et al. 2013b, 2014; Wang and Tang 2009; Zhou et al. 2012]. They use as a model a larger database of photo-style exemplar pairs (e.g., *CUHK Face Sketch Database* [Wang and Tang 2009] with 88 training and 100 testing faces), which can reproduce a much larger variety of target faces. Nevertheless, the data preparation phase for a different artistic style would be very tedious and time-consuming since the artist would need to prepare many photo-style exemplars for different subjects. This drawback can be partially alleviated by using techniques that understand the example-based process on the level of individual strokes [Berger et al. 2013; Wang et al. 2013a; Zhao and Zhu 2011]. However, although these methods achieve pleasing results for some styles of sketchy or painterly rendering, artistic styles that cannot be simply reduced to stroke-based decomposition are difficult to handle.

Recently, Gatys et al. [2016] demonstrated that guided parametric texture synthesis [Portilla and Simoncelli 2000] with an image representation based on deep neural networks [Simonyan and Zisserman 2014] can achieve impressive example-based style transfer. Advantages of this technique are that it requires only one exemplar image and that it can handle various different styles. Johnson et al. [2016] later provided a solution based on feed-forward networks that solves the original optimization problem much faster. Ruder et al. [2016] extended the original technique to handle video, and

Selim et al. [2016] provided improvements that give better results for portrait stylization, including portrait videos.

Despite the great success of neural-based style transfer techniques, a key limitation is the inability to faithfully reproduce low-level textural details (see Fig. 2). Fišer et al. [2016] proposed an alternative solution based on guided non-parametric texture synthesis that is able to preserve textural details, however, they use guidance channels that are tailored to stylized renderings of 3D models. This makes their approach inapplicable to portraits without modification.

Another disadvantage of existing neural-based video synthesis methods [Ruder et al. 2016; Selim et al. 2016] is that their aim is to achieve full temporal coherence. This conflicts with the temporal properties of real hand-crafted animations, in which a certain amount of temporal noise is always visible [Noris et al. 2011]. The Color Me Noisy method proposed by Fišer et al. [2014] allows introduction of temporal noise into an existing sequence using a randomized variant of a hierarchical texture synthesis algorithm [Wexler et al. 2007]. They initialize the synthesis with a sub-sampled version of the target frame and synthesize the remaining high-frequency details. A key drawback of this method is that the low-frequency content of the output sequence needs to be known beforehand. Also the texture synthesis method of Wexler et al. often fails to preserve textural richness of the style exemplar. Due to those limitations Color Me Noisy is not sufficient for our portrait stylization scenario.

Besides techniques that transfer traditional artistic media styles to head portraits, there are also approaches that transfer a specific photographic look [Kemelmacher-Shlizerman 2016; Shen et al. 2016; Shih et al. 2014; Yang et al. 2015]. Although their goal is not artistic style transfer, internally they use tools such as equalization of intensity levels and local contrast enhancement that can also be used in our domain.

Our approach also shares ideas with methods for novel view synthesis [Rematas et al. 2014] and constrained texture transfer [Diamanti et al. 2015; Jamriška et al. 2015; Kaspar et al. 2015; Lukáč et al. 2015, 2013; Ritter et al. 2006; Zhou et al. 2017] that further extend the original *texture-by-numbers* concept of Hertzmann et al. [2001]. However, none of these approaches provide a solution for faithful artistic style transfer for facial animations.

3 OUR APPROACH

The input to our method (Fig. 4) is a *style exemplar* image S of a stylized head portrait and a *target* video sequence T of a human facial performance. We assume the subject is mostly facing the camera and is not occluded by other objects (e.g., is not wearing glasses). The task is to produce a stylized sequence O that conveys the visual properties of the style S and respects the subject's facial characteristics so that the subject can be easily recognized from the stylized sequence. In addition we need O to follow the motion of T in a temporally coherent manner, while at the same time letting the user control the amount of temporal noise.

To solve this task we apply guided texture synthesis [Fišer et al. 2016], which has the ability to preserve fine textural details of the style exemplar. This approach is based on non-parametric texture synthesis [Kwatra et al. 2005; Wexler et al. 2007], which composes the target image by finding and blending suitable source patches. However, unlike standard texture synthesis, which uses only RGB values as guidance, the individual pixels in our source and target images contain additional guiding channels. These bias the selection of source patches towards a preferred subset more suitable for the particular semantic region in the target image (see, e.g., the texture-by-numbers application of Herzmann et al. [2001]).

Fig. 4. The goal of our method is to transfer the style from a single style exemplar (S) to a sequence of target frames T, producing the stylized output frames O. Style exemplar: © *Graciela Bombalova-Bogra*. Target frames: © *Ted Forbes* via YouTube.

Within the framework of guided texture synthesis our goal is to design a set of guiding channels tailored to head portrait videos, enabling rich, semantically meaningful style transfer with a controllable amount of temporal dynamics.

3.1 Overview

To produce compelling style transfer results, the guidance channels need to satisfy a few requirements.

Fišer et al. [2016] showed that artists typically use unique stylization for different semantic regions in the stylized scene. This applies to our domain as well; for example, in the painting in Fig. 3 the brush strokes in the forehead are much larger than those around the eyes. Motivated by this, we generate the *segmentation guide* G_{seg} (Fig. 3) that subdivides the head into hair, eyebrow, nose, lip, oral cavity, eye, and skin segments (see Section 3.2). To further encourage local consistency of the style transfer we introduce a *positional guide* G_{pos} that encourages the transfer of source patches to similar relative positions in the target (see Section 3.3).

Preserving basic shading cues maintains proper facial proportions and considerably helps the human visual system recognise the subject's identity [Sinha et al. 2006]. However, the overall appearance of the style exemplar and the target may differ considerably. To alleviate this difference, we remap intensity levels and local contrast values in the target image to be as close as possible to those in the style exemplar while still preserving the original shading cues. Such modified image is then used as an additional *appearance guide* G_{app} (see Section 3.4).

Finally, to produce visually pleasing video, there needs to be temporal guidance controllable by the user. In our scenario we try to preserve the appearance of hand-drawn sequences, which exhibit a certain amount of temporal flickering. To simulate such a phenomenon we combine the approach of LazyFluids [Jamriška et al. 2015] with an idea from Color Me Noisy [Fišer et al. 2014] and introduce a *temporal guide* G_{temp} (see Section 3.5).

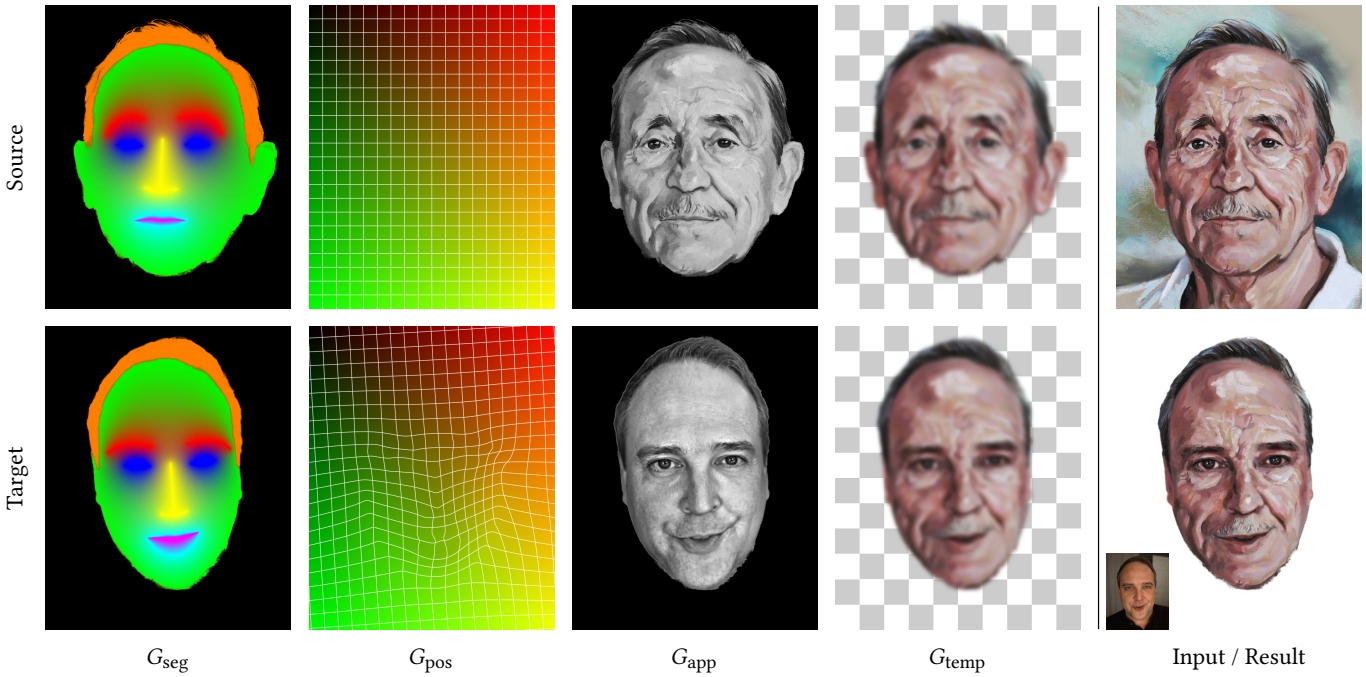

Fig. 3. Overview of the guiding channels used to perform the synthesis. The segmentation guide G_{seg} subdivides the source and target faces into a set of semantically meaningful regions (hair, eyebrows, nose, lips, oral cavity, eyes, and skin), for which an artist typically uses a specific stylization. The positional guide G_{pos} encourages the source patches to be transferred to similar relative positions in the target image. The appearance guide G_{app} helps enhance the perception of subject's identity by preserving local shading gradients in the target image. The temporal guide G_{temp} contains blurred versions of the source exemplar and of a motion-warped version of the previous frame to ensure coherency in the temporal domain. The amount of blur controls the amount of temporal flickering in the output sequence. The style exemplar and the resulting output with the original target frame are on right. Style exemplar: © *Graciela Bombalova-Bogra*. Target frame: © *Ted Forbes* via YouTube.

3.2 Segmentation guide

To generate the segmentation guide G_{seg} of a target frame T_i (Fig. 5a) we evaluated current state-of-the-art neural-based techniques [Jackson et al. 2016; Liu et al. 2015], but found them insufficiently accurate for our purpose (see Fig. 6). Instead we use a different approach that creates soft masks for the whole head and the skin region using closed-form matting [Levin et al. 2008]. It takes as input a coarse trimap that categorizes pixels as being definitely inside the region, definitely outside, or uncertain.

To create the head region trimap (Fig. 5b) we first erode and dilate a foreground mask obtained from automatic portrait segmentation [Shen et al. 2016] (Fig. 5c). This step helps to separate pixels that are assumed to be definitely inside and outside the head region. To detach the face region from the neck we further refine the trimap using a detected chin landmark [Kazemi and Sullivan 2014] (Fig. 5d). We render this landmark as a thick line of uncertain pixels and mark the disconnected neck as being definitely outside. Finally we apply closed-form matting to obtain the resulting soft mask (Fig. 5e).

To construct the skin region trimap (Fig. 5f), we use a simple statistical model of the skin. As observed by Gong and Sakauchi [1995], separating the chromatic and luminance components helps the segmentation of human skin. Therefore, we convert the image to YC_BC_R color space and fit the histogram of C_B and C_R components of cheek pixels with a multivariate Gaussian distribution. With it, we can determine the likelihood of each pixel being a skin pixel (Fig. 5g). We then normalize the likelihood map and consider all pixels above 0.5 to be part of the skin and adjust its trimap estimate (Fig. 5f) from which we generate the soft mask for the skin region (Fig. 5h).

The pixelwise difference of the facial and skin masks effectively segment the hair region (the orange color in Fig. 5i). Masks for the remaining segments – eyes, lips, oral cavity, nose, and eyebrows – are estimated using detected facial landmarks (Fig. 5d). Since the position of landmarks may be inaccurate, we avoid hard transitions by blurring the segment boundaries using diffusion curves [Orzan et al. 2008] (Fig. 5i).

A similar approach can also be applied to create the segmentation guide G_{seg} for the style exemplar S. However, in this case the style image can notably differ from the appearance of a real human, so parts of this automatic pipeline such as landmark and skin detection might fail. In this case the user helps the system by correcting trimaps and specifying better positions for landmarks. While this is additional manual intervention, it only needs to be done once and can be reused for the whole sequence.

3.3 Positional guide

The positional guide G_{pos} for the style exemplar is very simple; each pixel encodes its (x, y) coordinates normalized to the range of 0–1.

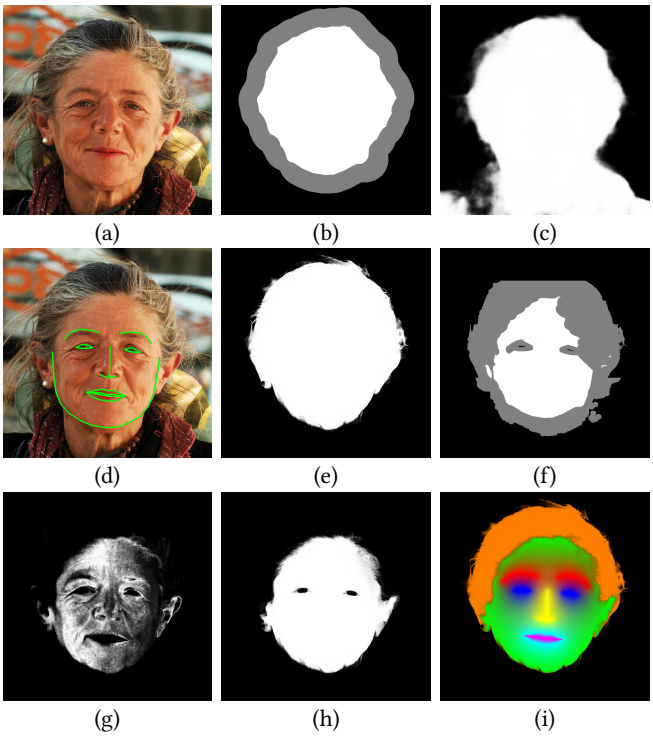

Fig. 5. Generating the segmentation guide G_{seg}. Given an input image (a), we obtain a head trimap (b) using an initial portrait segmentation [Shen et al. 2016] (c) and detected facial landmarks [Kazemi and Sullivan 2014] (d). A soft mask of the head is computed from the head trimap using closed-form matting [Levin et al. 2008] (e). A skin trimap (f) is estimated by thresholding the per-pixel likelihood of being a skin pixel (g). We compute a soft mask (h) from the resulting skin trimap. To obtain the remaining segments we use diffusion curves [Orzan et al. 2008] seeded with the facial landmarks, and subtract the skin from the head mask to obtain the hair region shown in orange (i). Input image © *Pedro Ribeiro Simões* via flickr.

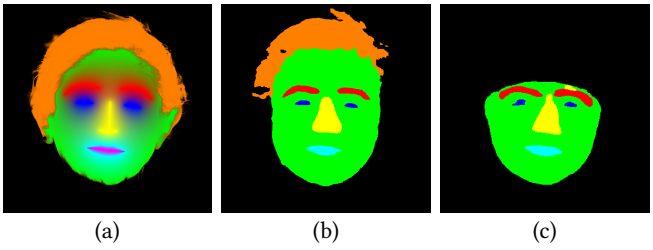

Fig. 6. Comparison of facial segmentations produced by our approach (a) with current neural-based state-of-the-art: Liu et al. [2015] suffers from inaccurately shaped segments (b) while the method of Jackson et al. [2016] does not support a hair segment and produces errors like the nose segment above the left eyebrow (c).

To generate G_{pos} for the target image, we use the detected facial landmarks in the style image and the corresponding ones in the target frame. We warp the exemplar's G_{pos} image using moving least squares deformation [Schaefer et al. 2006], where positions of the facial landmarks in the target image T_i and their connections are used as control lines to specify constraints for the resulting deformation field.

3.4 Appearance guide

To generate the appearance guide G_{app}, we convert the target image T_i and the style exemplar S to grayscale. Then we use the method of Shih et al. [2014] to modify the global intensity levels and local contrast values of the target image T_i to match those in the style exemplar S. To balance the tradeoff between preserving the subject's identity and retaining the textural richness of the style exemplar, we add an additional weighting channel that boosts the influence of G_{app} at certain pixels. Our experiments showed that the eyes and oral cavity regions need to have appearance closer to the target image and thus we use higher weights for the appearance guide to deliver more convincing stylization results. Section 3.6 discusses further refinements to the eye and mouth synthesis.

Our weighting scheme also gives the user additional artistic control to obtain a smooth transition between the identity of the subject in the target image and in the style exemplar. A higher weight for G_{app} makes the results look closer to the target image (Fig. 11).

3.5 Temporal guide

For full temporal coherence, we could have applied the approach of Jamriška et al. [2015], in which the previously synthesized frame O_{t-1} is advected by the underlying motion field (we estimate it using SIFT flow [Liu et al. 2011]) and used as a guide for the synthesis of a new frame. However, since we would like to preserve the appearance of hand-drawn sequences, which exhibit a certain amount of temporal dynamics, we also take into account an observation made by Fišer et al. [2014], that in real hand-drawn sequences the temporal coherence is preserved only at lower frequencies.

In the Color Me Noisy scenario, Fišer et al. assume that the low-frequency content of the source and target are the same, allowing the synthesis to be started at a certain resolution level. This is, however, not satisfied in our scenario since the style exemplar can differ significantly from the target. Instead we propose a different solution that follows the Color Me Noisy principle to preserve the temporal coherence at lower frequencies, but does not require the source and target to match. We blur the style exemplar S and the previously synthesized frame O_{t-1} after advection, and use them as a *temporal guide* G_{temp}. The amount of temporal flickering is then controlled by varying the cut-off frequency (width) of the blurring kernel. Another advantage of this solution is that it decouples control over the amount of temporal noise from other aspects of the synthesis as will be demonstrated in Section 4.

3.6 Special treatment of open mouth and eyes

Special handling is required when the style exemplar has a closed mouth, but the target frame shows the subject's teeth. The guidance channels described so far bias the synthesis towards using lip patches for the teeth, leading to an unnatural and unappealing result (Fig. 7b). To improve the quality we create a special mask with increased weight of G_{app} using the detected facial landmarks (Fig. 7d). This map allows us to guide the synthesis to transfer lighter

texture areas of the style exemplar to the teeth, even if they are distant, creating a more plausible result (Fig. 7c).

Our experimentation showed that even minor defects in the eye stylization lead to disturbing results, and without modification the eyes often do not resemble the style exemplar (Fig. 7b). To address this we synthesize the eyes separately, based on a special set of guiding channels with only a hard segmentation and a normal map (Fig. 7e, f, g). We construct these channels using the method of Johnston et al. [2002]. After synthesizing the face, we blend in the synthesized eyes using a soft mask of the target head to produce the final output (Fig. 7c).

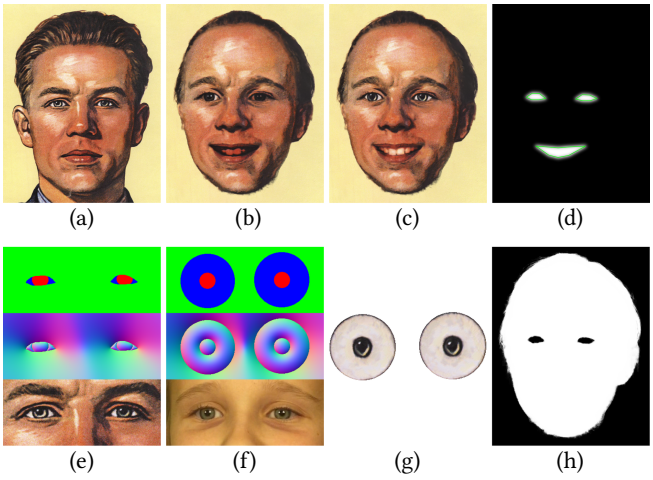

Fig. 7. Special handling of eyes and open mouth. Without special care, the synthesis of an open mouth does not look appealing (b). To improve that, a mask with increased weight of G_{app} (d) is used to guide the transfer of lighter texture areas of the style exemplar to the teeth (c). Special handling for the eyes is also needed to preserve the appearance of the style exemplar and avoid disturbing results (b). We synthesize the eyes separately, using a hard segmentation (iris vs. sclera) and a normal map as guiding channels, computed using [Johnston 2002] (e, f). The synthesized eyes (g) are then composited with the rest of the face using soft mask of the target head (h) to produce the final output (c). Style exemplar: Viktor Ivanovich Govorkov. Target subject: © *Štěpánka Sýkorová.*

3.7 Synthesis

Once we have the guiding channels (G_{seg}, G_{pos}, G_{app}, G_{temp}) we can run the guided texture synthesis algorithm of Fišer et al. [2016]. A key advantage of this technique is that it adaptively encourages uniform utilization of source patches and thus suppresses the "wash-out" effect [Jamriška et al. 2015] inherent to other texture synthesis techniques based on the original texture optimization strategy [Kwatra et al. 2005; Wexler et al. 2007].

In our solution we also need to address a "floating texture" artifact described by Fišer et al. [2014], which is the formation of distracting coherent islands of patches that become visible when a sequence of images produced by non-parametric texture synthesis is played back. To break those islands for every stylized frame Fišer et al. modify the style exemplar using a randomized free-form deformation. This change guarantees that the newly synthesized frame cannot contain

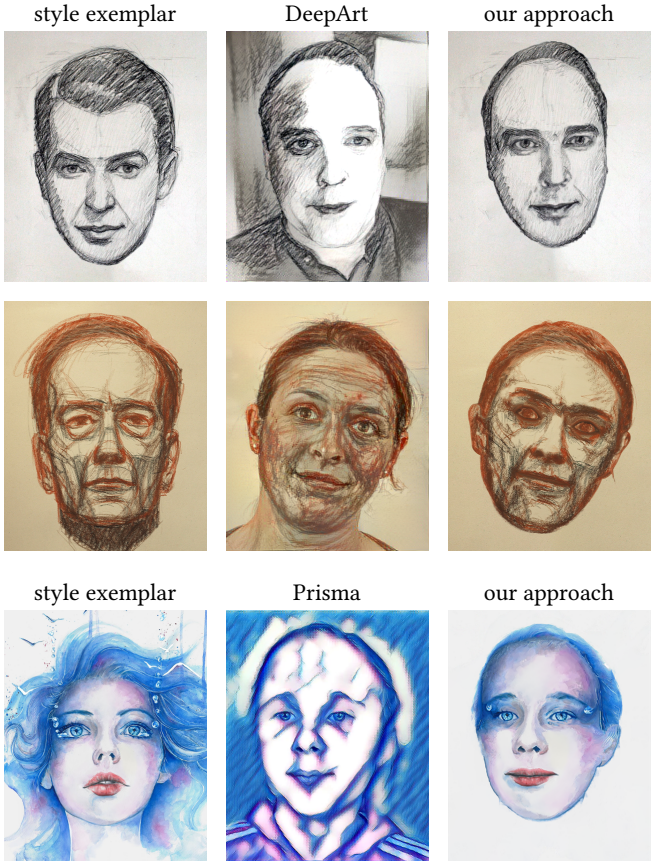

Fig. 12. Comparison with publicly available implementations of neural-based style transfer to video: *DeepArt* web service based on Gatys et al. [2016] & Ruder et al. [2016] and *Prisma* mobile application inspired by the method of Johnson et al. [2016]. Style exemplars (top to bottom): © *Adrian Morgan* via flickr, © *Adrian Morgan* via flickr, © *Joanna Wedrychowska* via instagram. Target subjects (top to bottom): *Ted Forbes*, *Gabriela Daniels*, *Štěpánka Sýkorová.*

the same static region of pixels as the previous frame. However, a fundamental issue here is that free-form deformation in fact breaks the low-level textural consistency of the used artistic media. To alleviate this drawback in our solution we only slightly rotate the style exemplar to match the dominant rotation of the subject's face in the target sequence. This makes the change of style exemplar consistent with the global orientation difference between the style exemplar and target patches. We estimate the closest relative rotation that aligns corresponding source and target chin landmarks to have a minimal distance in the least squares sense using the closed-form solution described in Schaefer et al. [2006].

4 RESULTS

We implemented our method using C++ and CUDA. On a 3 GHz quad-core CPU it takes around 30 seconds to compute all necessary guiding channels for a one-megapixel frame. For the subsequent synthesis we use 5 pyramid levels. On each resolution level we run

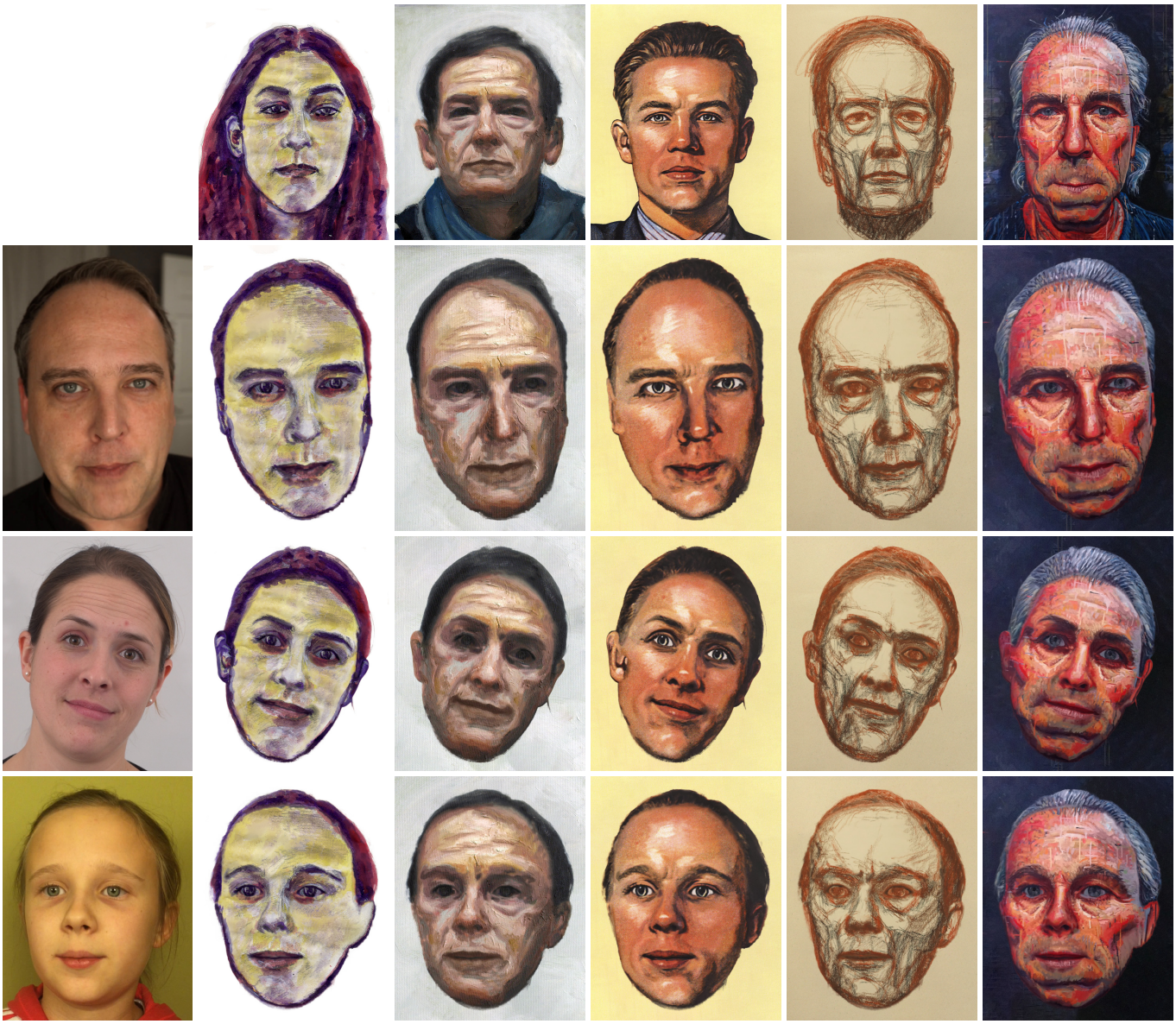

Fig. 8. Results. The style exemplars across the top have been applied to the subjects in the left column. Target subjects (top to bottom): © *Ted Forbes* via YouTube, © *Gabriela Daniels*, © *Štěpánka Sýkorová*. Style exemplars (left to right): © *Arturo Espinosa* via flickr, © *Adrian Morgan* via flickr, Viktor Ivanovich Govorkov, © *Adrian Morgan* via flickr, © *Matthew Cherry* via http://matthewivancherry.com/home.html and https://www.instagram.com/matthewivancherry.artist (HAT, oil on canvas, 48" x 48", 2011).

4 voting iterations with 4 PatchMatch sweeps [Barnes et al. 2009]. PatchMatch is executed selectively, only done for those patches that improved in the previous step. Also for every guiding channel in each patch we measure the standard deviation σ of pixel values. When $\sigma < 0.01$ we approximate the error metric using only the single squared difference of mean values instead of the w^2 squared differences normally used for a patch of width w. With these approximations we can synthesize a one-megapixel frame in 3 minutes on the CPU and in 5 seconds on the GPU (GeForce GTX 970).

For the previously published methods upon on which our pipeline is built we set parameter values as recommended in the corresponding papers. We also fine-tuned the specific weights for the individual guiding channels. G_{seg} and G_{pos} have weight 5. The appearance channel G_{app} has weight 1 except in the eye and mouth regions, where it is set to 5. Channel G_{temp} and the style channel have their weights set to 3. Those values were used to generate all results shown in this paper.

Fig. 9. Additional results. A variety of style exemplars have been applied to the male subject from Fig. 8. Target subject: © *Ted Forbes* via YouTube. Style exemplars (left to right): © *Adrian Morgan* via flickr, © *Thomas Shahan* via flickr, *Egon Schiele*, © *Arturo Espinosa* via flickr, © *Scary Zara Mary* via facebook, Kazimir Malevich.

Fig. 10. Necessity of individual guiding channels. Without the segmentation guide G_{seg}, exemplar patches are used in locations that are not semantically meaningful. Without the positional guide G_{pos}, semantically meaningful patches are used at distant locations, e.g., patches from the sideburns are used on top of the head. Without the appearance guide G_{app}, the target subject's identity is not preserved, e.g. the nose width and eye size. Target subject: © *Jakub Fišer*. Exemplar style: © *Matthew Cherry* via http://matthewivancherry.com/home.html and https://www.instagram.com/matthewivancherry.artist (HAT, oil on canvas, 48" x 48", 2011).

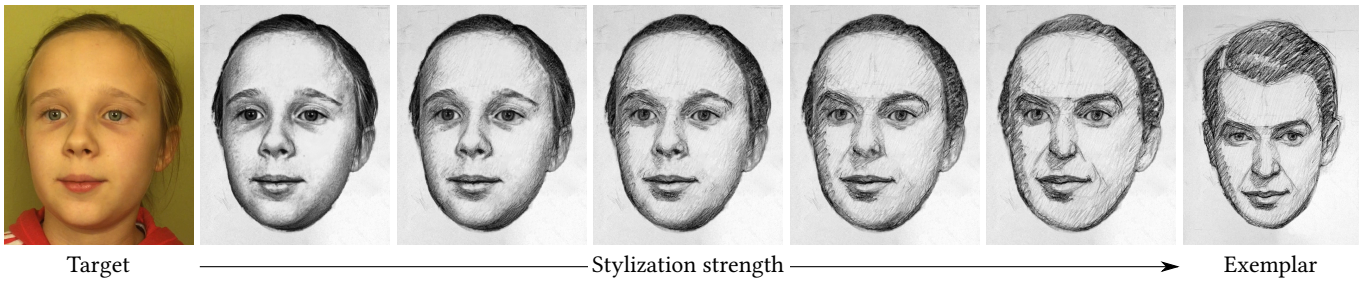

Fig. 11. Stylization strength. Controlling the weight of the appearance guide G_{app} allows a smooth transition between the identity of the subject in the target sequence (left) and in the style exemplar (right). Target subject: © *Štěpánka Sýkorová*. Exemplar style: © *Adrian Morgan* via flickr.

Fig. 13. Comparison with neural-based style transfer [Selim et al. 2016]. Additional comparisons are in Fig. 2. Style exemplars (top to bottom): © *Graciela Bombalova-Bogra*, © *Scary Zara Mary* via facebook, © *Jen Garcia* via flickr. Target subjects (top to bottom): *Nick P. Law*, *Barack Obama*, *Anne Hathaway*.

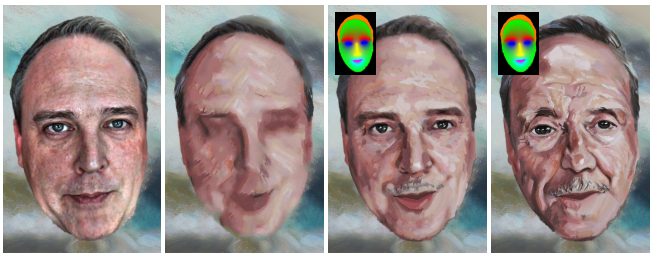

Fig. 14. Based only on color information transferred from the style exemplar to the target frame [Shih et al. 2014] (a), Color Me Noisy [Fišer et al. 2014] is unable to synthesize an accurate result (b). When provided with our segmentation guide G_{seg}, the results improve notably, however, with low temporal noise the synthesis suffers from the "wash-out" effect [Jamriška et al. 2015] (c), and with high temporal noise, fidelity to the target is lost (d). The style exemplar is shown in Fig. 13. Target subject: © *Ted Forbes* via YouTube.

To validate the need for all the guiding channels, we show the synthesis results when individual channels have been selectively turned off (Fig. 10). Without G_{seg} the style is transferred to improper locations in the target, without G_{pos} an unnatural mixture of patches destroys the overall appearance of the style exemplar, and without G_{app} the overall appearance of the target's subject is lost. Finally, eliminating G_{temp} removes temporal coherence in the video result.

As shown in Fig. 10, the appearance guide G_{app} has a major impact on the perceived subject's identity. We set this guiding channel to have greater weight in visually important regions (eyes, mouth, and their neighbourhood).

Fig. 11 shows how controlling the weight of the appearance guide G_{app} allows a smooth transition between matching the target and preserving the source style.

We tested our approach on various head portrait sequences of men, women, and children using different styles and artistic media, including watercolor, oil paint, pastel, pencil, engraving, and acrylic paint (Figures 1, 8, and 9). In contrast to neural-based style transfer [Gatys et al. 2016; Johnson et al. 2016; Selim et al. 2016], which has difficutlies to preserve low-level textural details of the transferred style our technique retains visual richness for a wide variety of styles (Figures 2, 12, and 13). Moreover, the controllable amount of temporal flickering visible in the supplementary video notably enhances the perception of authentic hand-painted content, in contrast to previous fully temporally coherent attempts [Ruder et al. 2016; Selim et al. 2016].

We also compared our approach with the Color Me Noisy technique of Fišer et al. [2014]. To get closer to their original assumption that the target sequence has similar appearance as the style exemplar, we applied the method of Shih et al. [2014] to make the appearance of the target match that of the style exemplar (Fig. 14a). Video results clearly show that the Color Me Noisy approach introduces temporal flickering comparable to our approach. However, lack of guidance leads to patches being transferred inappropriately (Fig. 14b). Results can be improved if guidance is added, but at low temporal noise levels, low-level textural details are lost leading to the "wash-out" effect [Jamriška et al. 2015] (Fig. 14c). This is caused by Color Me Noisy using the method of Wexler at al. [Wexler et al. 2007] to perform the synthesis. At high temporal noise levels, the identity of the target subject is lost (Fig. 14d). Our approach decouples these effects. It controls the amount of visual noise separately from the resemblance between the identity of the subject in the target sequence and the style exemplar.

Besides artistic style exemplars we also experimented with realistic exemplars such as statues (Fig. 15) and portrait photographs (Fig. 16). The results show that our method has the potential to become a more general appearance transfer tool. We envision further development in this direction in the future.

5 LIMITATIONS AND FUTURE WORK

Although our stylization approach produces compelling results for a variety of input styles and target subjects, it still has some limitations.

Our approach has problems with style exemplars that contain longer semantically important linear structures, like the forehead wrinkles in Egon Schiele's style exemplar in the third column of Fig. 9

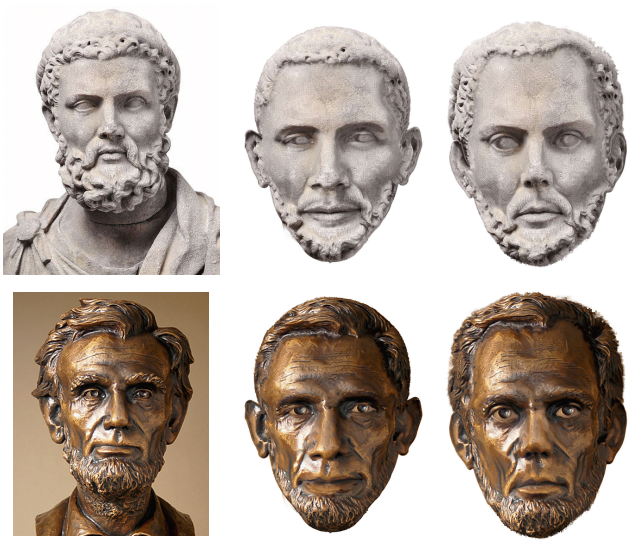

Fig. 15. Using our method to do photorealistic style transfer. A photograph of a statue is used as a style exemplar (left) and transferred to two different subjects (right). Style exemplars (top to bottom): © *Country French Interiors*, © *Will Murray*. Target subjects (left to right): *Barack Obama*, *Nick P. Law*.

and the corresponding synthesis result. To alleviate these problems some additional user guidance could be provided in the spirit of the geometric constraints used in the PatchMatch algorithm [Barnes et al. 2009].

Another limiting factor is the relatively restricted pool of available exemplar patches, caused using a single style image for the synthesis of the entire sequence. In real hand-painted animations, fresh new strokes appear with every frame, making the content less repetitive and visually more attractive. Such a drawback can be alleviated using multiple exemplar images at the expense of taking more time for the style preparation phase. One could also incorporate an incremental approach proposed by Bénard et al. [2013], in which the artist locally modifies already-synthesized parts of the animation and has those changes immediately propagated to other frames.

Finally, a drawback of both our technique and other recent style-transfer approaches is that they predominantly modify the subject's color and texture while leaving the subject's global shape characteristics unchanged. In real paintings, however, appearance and shape stylization are coupled together to jointly represent a certain artistic expression. Our failure to accommodate this can cause a notable difference between the shape of the face in the style exemplar and that in the synthesized image. In future work we plan to explore how to incorporate shape aspects of the style exemplar.

A related issue, which we left for future work, is the separation of style (e.g., the size of the brush strokes) and content (e.g., wrinkles or a mustache). It is hard to decide algorithmically which components of the exemplar should be transferred.

6 CONCLUSION

We have presented an approach to example-based stylization of head portrait videos. Our technique automatically transfers style from a single exemplar image to a head portrait in motion while preserving the subject's identity. It transfers both the overall look and the low-level textural details of the exemplar. We introduce a controllable amount of temporal flickering, which creates the perception that the sequence was painted frame-by-frame. Despite the current trend in applying neural-based techniques to image synthesis tasks, our results confirm that even a simple non-parametric texture synthesis framework can achieve state-of-the-art results.

Fig. 16. An example of inverse stylization. We take a photo of a real subject (left) and apply it as a "style" to a sketch (middle). The resulting image (right) resembles the original photo while retaining the overall structure of the sketch. Target subject: © *Profimedia*. Exemplar style: © *Jared Houghton* via DeviantArt.

ACKNOWLEDGEMENTS

We would like to thank numerous artists for providing style exemplars, our target subjects for their videos, Manuel Ruder & Ahmed Selim for help with comparison, Vojtěch Votýpka for recording the video of Gabriela Daniels, and all anonymous reviewers for insightful comments and suggestions. This research began as an internship by Jakub Fišer at Adobe. It was funded by Adobe and has been supported by the Technology Agency of the Czech Republic under research program TE01020415 (V3C – Visual Computing Competence Center) and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS16/237/OHK3/3T/13 (Research of Modern Computer Graphics Methods). Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

REFERENCES

Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics* 28, 3 (2009), 24.

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.

Itamar Berger, Ariel Shamir, Moshe Mahler, Elizabeth J. Carter, and Jessica K. Hodgins. 2013. Style and abstraction in portrait sketching. *ACM Transactions on Graphics* 32, 4 (2013), 55.

Hong Chen, Lin Liang, Ying-Qing Xu, Heung-Yeung Shum, and Nan-Ning Zheng. 2002. Example-Based Automatic Portraiture. In *Proceedings of Asian Conference on Computer Vision*. 171–178.

Hong Chen, Ziqiang Liu, Chuck Rose, Yingqing Xu, Heung-Yeung Shum, and David Salesin. 2004. Example-Based Composite Sketching of Human Portraits. In *Proceedings of International Symposium on Non-photorealistic Animation and Rendering*.

95–102.
Hong Chen, Nanning Zheng, Lin Liang, Yan Li, Ying-Qing Xu, and Heung-Yeung Shum. 2002. PicToon: A Personalized Image-Based Cartoon System. In *Proceedings of ACM International Conference on Multimedia*. 171–178.
Doug DeCarlo and Anthony Santella. 2002. Stylization and Abstraction of Photographs. *ACM Transactions on Graphics* 21, 3 (2002), 769–776.
Olga Diamanti, Connelly Barnes, Sylvain Paris, Eli Shechtman, and Olga Sorkine-Hornung. 2015. Synthesis of Complex Image Appearance from Limited Exemplars. *ACM Transactions on Graphics* 34, 2 (2015), 22.
Steve DiPaola. 2007. Painterly rendered portraits from photographs using a knowledge-based approach. In *Proceedings of SPIE Human Vision and Electronic Imaging*, Vol. 6492. 33–43.
Alexei A. Efros and William T. Freeman. 2001. Image Quilting for Texture Synthesis and Transfer. In *SIGGRAPH Conference Proceedings*. 341–346.
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, 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.
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.
Yihong Gong and Masao Sakauchi. 1995. Detection of Regions Matching Specified Chromatic Features. *Computer Vision and Image Understanding* 61, 2 (1995), 263–269.
Bruce Gooch, Erik Reinhard, and Amy Gooch. 2004. Human Facial Illustrations: Creation and Psychophysical Evaluation. *ACM Transactions on Graphics* 23, 1 (2004), 27–44.
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.
Aaron Jackson, Michel Valstar, and Georgios Tzimiropoulos. 2016. A CNN Cascade for Landmark Guided Semantic Part Segmentation. In *Proceedings of ECCV 2016 Workshops, Geometry meets Deep Learning*.
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.
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.
Scott F. Johnston. 2002. Lumo: Illumination for Cel Animation. In *Proceedings of the 2Nd International Symposium on Non-photorealistic Animation and Rendering*. 45–ff.
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.
Ira Kemelmacher-Shlizerman. 2016. Transfiguring Portraits. *ACM Transactions on Graphics* 35, 4 (2016), 94.
Vivek Kwatra, Irfan A. Essa, Aaron F. Bobick, and Nipun Kwatra. 2005. Texture optimization for example-based synthesis. *ACM Transactions on Graphics* 24, 3 (2005), 795–802.
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.
Anat Levin, Dani Lischinski, and Yair Weiss. 2008. A Closed-Form Solution to Natural Image Matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 2 (2008), 228–242.
Hongliang Li, Guanghui Liu, and King Ngi Ngan. 2011. Guided Face Cartoon Synthesis. *IEEE Transactions on Multimedia* 13, 6 (2011), 1230–1239.
Ce Liu, Jenny Yuen, and Antonio Torralba. 2011. SIFT Flow: Dense Correspondence across Scenes and Its Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 5 (2011), 978–994.
Sifei Liu, Jimei Yang, Chang Huang, and Ming-Hsuan Yang. 2015. Multi-Objective Convolutional Learning for Face Labeling. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
Michal Lukáč, Jakub Fišer, Paul Asente, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. 2015. Brushables: Example-based Edge-aware Directional Texture Painting. *Computer Graphics Forum* 34, 7 (2015), 257–268.
Michal Lukáč, Jakub Fišer, Jean-Charles Bazin, Ondřej Jamriška, Alexander Sorkine-Hornung, and Daniel Sýkora. 2013. Painting by Feature: Texture Boundaries for Example-based Image Creation. *ACM Transaction on Graphics* 32, 4 (2013), 116.
Meng Meng, Mingtian Zhao, and Song Chun Zhu. 2010. Artistic paper-cut of human portraits. In *Proceedings of ACM Multimedia*. 931–934.
Gioacchino Noris, Daniel Sýkora, Stelian Coros, Brian Whited, Maryann Simmons, Alexander Hornung, Marcus Gross, and Robert Sumner. 2011. Temporal Noise Control for Sketchy Animation. In *Proceedings of International Symposium on Non-photorealistic Animation and Rendering*. 93–98.
Alexandrina Orzan, Adrien Bousseau, Holger Winnemöller, Pascal Barla, Joëlle Thollot, and David Salesin. 2008. Diffusion Curves: A Vector Representation for Smooth-Shaded Images. *ACM Transactions on Graphics* 27, 3 (2008), 92.
Javier Portilla and Eero P. Simoncelli. 2000. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision* 40, 1 (2000), 49–70.
Konstantinos Rematas, Tobias Ritschel, Mario Fritz, and Tinne Tuytelaars. 2014. Image-Based Synthesis and Re-synthesis of Viewpoints Guided by 3D Models.. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 3898–3905.
Lincoln Ritter, Wilmot Li, Brian Curless, Maneesh Agrawala, and David Salesin. 2006. Painting With Texture. In *Proceedings of Eurographics Symposium on Rendering*. 371–376.
Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. 2016. Artistic Style Transfer for Videos. In *Proceedings of German Conference Pattern Recognition*. 26–36.
Scott Schaefer, Travis McPhail, and Joe Warren. 2006. Image Deformation Using Moving Least Squares. *ACM Transactions on Graphics* 25, 3 (2006), 533–540.
Ahmed Selim, Mohamed Elgharib, and Linda Doyle. 2016. Painting Style Transfer for Head Portraits Using Convolutional Neural Networks. *ACM Transactions on Graphics* 35, 4 (2016), 129.
Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian L. Price, Eli Shechtman, and Ian Sachs. 2016. Automatic Portrait Segmentation for Image Stylization. *Computer Graphics Forum* 35, 2 (2016), 93–102.
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.
Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).
Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. 2006. Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proc. IEEE* 94, 11 (2006), 1948–1962.
Patrick Tresset and Frédéric F. Leymarie. 2005. Generative Portrait Sketching. In *Proceedings of International Conference on Virtual Systems and Multimedia*. 739–748.
Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. 2013b. Transductive Face Sketch-Photo Synthesis. *IEEE Transactions on Neural Networks and Learning Systems* 24, 9 (2013), 1364–1376.
Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. 2014. A Comprehensive Survey to Face Hallucination. *International Journal of Computer Vision* 106, 1 (2014), 9–30.
Tinghuai Wang, John P. Collomosse, Andrew Hunter, and Darryl Greig. 2013a. Learnable Stroke Models for Example-based Portrait Painting. In *Proceedings of British Machine Vision Conference*.
Xiaogang Wang and Xiaoou Tang. 2009. Face Photo-Sketch Synthesis and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 11 (2009), 1955–1967.
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.
Holger Winnemöller, Sven C. Olsen, and Bruce Gooch. 2006. Real-time video abstraction. *ACM Transactions on Graphics* 25, 3 (2006), 1221–1226.
Ming Yang, Shu Lin, Ping Luo, Liang Lin, and Hongyang Chao. 2010. Semantics-driven portrait cartoon stylization. In *Proceedings of International Conference on Image Processing*. 1805–1808.
Yue Yang, Hanli Zhao, Lihua You, Renlong Tu, Xueyi Wu, and Xiaogang Jin. 2015. Semantic portrait color transfer with internet images. *Multimedia Tools and Applications* (2015), 1–19.
Kun Zeng, Mingtian Zhao, Caiming Xiong, and Song-Chun Zhu. 2009. From image parsing to painterly rendering. *ACM Transactions on Graphics* 29, 1 (2009), 2.
Yong Zhang, Weiming Dong, Oliver Deussen, Feiyue Huang, Ke Li, and Bao-Gang Hu. 2014. Data-driven Face Cartoon Stylization. In *SIGGRAPH Asia Technical Briefs*. 14.
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.
Hao Zhou, Zhanghui Kuang, and Kwan-Yee Kenneth Wong. 2012. Markov Weight Fields for face sketch synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 1091–1097.
Yang Zhou, Huajie Shi, Dani Lischinski, Minglun Gong, Johannes Kopf, and Hui Huang. 2017. Analysis and Controlled Synthesis of Inhomogeneous Textures. *Computer Graphics Forum* 36, 2 (2017).