Learning Selfie-Friendly Abstraction from Artistic Style Images*

Yicun Liu LIUYICUN@SENSETIME.COM
Jimmy Ren RENSIJIE@SENSETIME.COM
Jianbo Liu LIUJIANBO@SENSETIME.COM
Jiawei Zhang ZHANGJIAWEI@SENSETIME.COM
Xiaohao Chen CHENXIAOHAO@SENSETIME.COM
SenseTime Research

Editors: Jun Zhu and Ichiro Takeuchi

Abstract

Artistic style transfer can be thought as a process to generate different versions of abstraction of the original image. However, most of the artistic style transfer operators are not optimized for human faces thus mainly suffers from two undesirable features when applying them to selfies. First, the edges of human faces may unpleasantly deviate from the ones in the original image. Second, the skin color is far from faithful to the original one which is usually problematic in producing quality selfies. In this paper, we take a different approach and formulate this abstraction process as a gradient domain learning problem. We aim to learn a type of abstraction which not only achieves the specified artistic style but also circumvents the two aforementioned drawbacks thus highly applicable to selfie photography. We also show that our method can be directly generalized to videos with high inter-frame consistency. Our method is also robust to non-selfie images, and the generalization to various kinds of real-life scenes is discussed. We will make our code publicly available.

Keywords: Artistic Style Abstraction, Gradient Domain Learning, Computational Photography, Computer Vision

1. Introduction

In the art world of painting creation, realistic traits of real-world scenes are often represented by a variety of artistic abstractions. The traditional filtering method of image abstraction usually smoothed the image while preserving various levels of structure. Recently, with the rise of deep learning, a seminal method for style generation has been proposed by Gatys et al. (2016b).

So far, despite various methods of image abstraction and style generation have been explored, precisely depicting human faces in artistic style remains challenging due to the strait restrictions on structural realism and color consistency. On the one hand, the human visual system is incredibly sensitive to irregularities in faces Jing et al. (2017); Selim et al. (2016), even minor deformation in edges will affect the accuracy of human facial identification, leading to the unrealistic feeling of the poorly stylized version of human faces. On the

* Code available at: https://github.com/DandilionLau/Selfie-Friendly-Abstraction

Figure 1: Even if with color preservation constraints, many style generation algorithms still perform poorly on facial images. In our experiment, (a) is the original image, (b) is Fast Neural Style (Johnson et al., 2016) with luminance-only transfer (Gatys et al., 2016a) to preserve the original color, (c) is Deep Analogy Liao et al. (2017), and (d) is CNNMRF (Li and Wand, 2016). The style references for (c) and (d) is carefully selected similar female portrait paintings. Although these methods work fine in highly-abstracted style, they suffer from color shift and shape deformation when applied in fine-grained human facial cases. As comparison, image (e)(f)(g) are our results learned from different gradient-domain style.

other hand, because skin tone mainly serves as an essential visual feature for human faces, maintaining skin color in the stylistic version of selfies is crucial Shih et al. (2014).

In this paper, we aim to learn the 'selfie-friendly' abstraction to both precisely and vividly depict human faces in artistic styles. To this end, we proposed a selfie-optimized CNN with a gradient domain training procedure. Unlike previous schemes, our method aims at learning the style abstraction directly on the gradient domain of images and can well tackle the two aforementioned drawbacks. Our framework is capable of learning different artistic abstractions while preserving the structural and color information in the original selfie. Another benefit of this innovation is that the framework can be directly used to render artistic style videos, with no flicking effect and convincingly high inter-frame consistency.

To ensure the extent of style abstraction in sophisticated cases, we are the first to show that using the perceptual loss in the gradient domain can capture the stylistic traits of artistic images. Furthermore, the application of our framework is not limited to the selfie images. Thanks to the nature of gradient processing, it can also be generalized to diverse real-life scenarios which require color realism and structural consistency.

The main contribution of our work can be summarized in the following perspectives:

Figure 2: Selfie-friendly results generated by our algorithm: The first row are the original selfie photos. The second row are non selfie friendly references generated from existing method, which often come with severe distortion and color blemish. The third row are results generated from our method, learned from style sources in the second column. Please zoom in on monitor for better comparison.

- First, we investigate two critical drawbacks of previous style generation methods on human faces and propose our selfies-friendly style abstraction framework that fully circumvents these drawbacks and achieves more attractive results for facial image stylization.
- Second, we explore the potential of gradient domain learning in the task of style abstraction. Our novelty includes applying firstly perceptual loss directly on gradient domain and color recovery from gradient images to comprehensively retain the original skin color.
- Third, our method tackles the drawback of flicking effect in style videos. In the video stylization task, it manifests high inter-frame consistency and is capable of rendering flicking-free artistic style videos, which is highly applicable to daily video streaming.

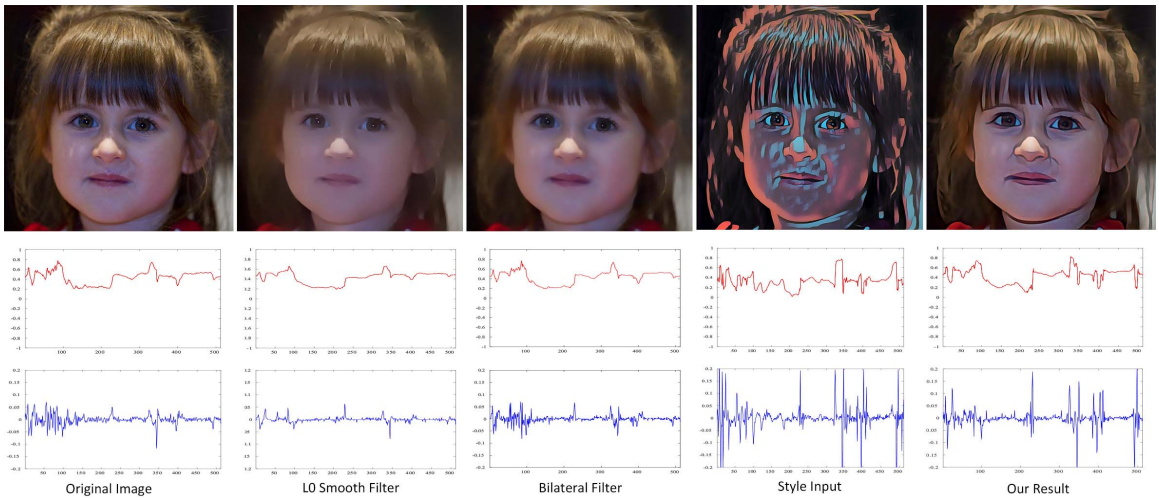

Figure 3: Color and gradient domain analysis: figures in the second row are the sectional views of the red channel, and figures in the third row are the sectional views of the horizontal gradient map. These two edge-aware filters smooth out most details except salient edges, with the variation being lower in both color and gradient domain. It is worth notice that the gradient view of our style input extremely diverts from the gradient views of the original image and edge-aware filter output, which is mainly resulted from distortions and color shift. Although in color domain, the structural information is usually mixed with chromatic information, the distinction is much more clear in gradient domain. In the training process, our method learns the edge-preserved style abstraction from candidate gradient patches. To this end, it is possible to diminish those defective scales and cracks and circumvent the visual drawbacks appearing in the style input.

2. Related Work

Artistic style abstraction and style generation have always been an open-ended challenge. Painting creation itself can be thought of a combination of the two tasks. Previously, image abstraction was investigated by many image processing papers. Many of the traditional filter-based methods handle image abstraction in an edge-aware manner, aiming to manipulate the rest of the image while preserving key structural information like edges. They have received a great deal of attention, like local Laplacian filter Paris et al. (2015), L0 smooth filter Xu et al. (2011), rolling guidance filter Zhang et al. (2014), and bilateral grid processing Chen et al. (2007). Those methods demonstrate satisfying results in image abstraction, but unable to deal with complex abstraction tasks like artistic style abstraction, which requires semantic information of the image rather than local filter processing.

As the other task, style generation was usually decomposed into multi-levels: brushstroke level Lu et al. (2010), texture level Efros and Freeman (2001) and patch level Meng et al. (2010). Recently, CNN based style generation algorithm like Gatys et al. (2016b); Johnson et al. (2016); Liao et al. (2017) aimed to learn the in-depth representation of style images, achieving impressive results. These techniques suffer from unwanted defects in facial depiction, ranging from deformation in facial edges to severe color shift in skin tone. Even with color preservation like luminance-only transfer Gatys et al. (2016a) to correct the skin color, there still exists severe blemish. An example is showed in figure 1.

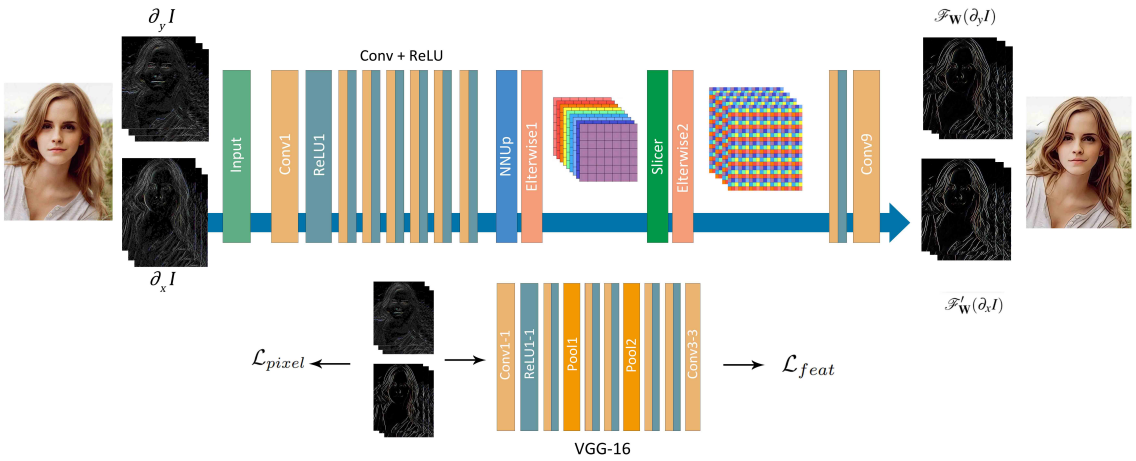

Figure 4: Architecture of our selfie-friendly stylization network. Our design operates on the gradient domain of input and reference images. Our network includes seven convolutional layers with ReLU to extract gradient features from the input images and the style references. Note that the height and width of the feature map in Conv-7 are downsampled by the scale factor of four. We later adopt the sub-pixel module to upsample the stylistic results to be the same as the original size. We calculate the pixel-wise loss $\mathcal{L}_{pixel}$ with the outputs of Conv-9 and style references. Then we used VGG-16 to calculate perceptual loss $\mathcal{L}_{feat}$ on the gradient domain.

Few published papers addressed the problem of unsatisfying defects in skin color and facial edges Selim et al. (2016); Sucontphunt (2014). Only limited amount of authors have noticed similar irregularities when they dealt with facial images. Their attempts included reducing portrait distortion by exploiting human face geometry Chen et al. (2001). and recovering skin smoothness from brushstroke transfer task O'Regan and Kokaram (2009). However, these approaches were used required intensive computational cost and were only applicable to limited styles like sketches style. There were also attempts to recover the original color in style transformation Gatys et al. (2017); Luan et al. (2017), with some methods using color histogram matching and luminance-only transfer Gatys et al. (2016a). Generally, their approaches suffered from extra tradeoff in the extent of stylization and were not specially designed for selfies. Many video-oriented stylization methods have been introduced Huang et al. (2017); Chen et al. (2017), but only a few were concentrating on maintaining inter-frame consistency in style transfer task has also been studied in Ruder et al. (2017).

The most proximate work to our goal is Selim et al. (2016); Shih et al. (2014). The algorithm proposed by Slim aimed to reduce the distortion and increase visual fidelity by adding an example-driven spatial constraint for each facial image. One limitation of the example-based method is that it requires similar portrait images as the reference, which is hard to find. Their model also lacked global consideration for skin color consistency. Unlike Selim et al. (2016), our method can be generalized to other relevant scenes and exhibits overall concern for multi-aspect naturalness preservation in the universal light-weighted framework.

Before our experiment, the topic of gradient domain processing has been merely explored since the rise of deep neural network Krizhevsky et al. (2012). The potential of using

gradient training in deep convolutional neural network has still not been throughout looked into. Early work mainly focused on gradient tone mapping and gradient dynamic range compression Socolinsky (2012). Xu is the first to employed gradient domain training to accelerate deep edge-aware filters, their results showed promising capabilities in preserving structural realism Xu et al. (2015). Gradient processing was also utilized in Mechrez et al. (2017) for photorealistic style transfer. Nevertheless, there is still no further examination for using gradient domain training in specific structure-color-concentrating tasks, such as style abstraction and generation for selfies.

3. Our Approach

Our method takes two images: an input image I which is usually an ordinary selfie image and a corresponding stylized reference of the original selfie $\mathcal{L}(I)$. $\mathcal{L}$ could be an existing style transfer algorithm, which usually generates reference stylized selfie images with apparent drawbacks of inaccurate edges and color shift. In this algorithm, we seek the optimized facial transformation $\mathcal{F}$ which transfers selfie image I to a visually more satisfying stylized version $\mathcal{F}_{\mathbf{W}}(I)$. Here $\mathcal{F}$ denotes the network architecture, and $\mathbf{W}$ represents the network parameters. We name this optimized transformation as selfie-friendly transformation, which should not only generate the artistic abstraction of the reference but also avoid the aforementioned drawbacks.

Our approach achieves selfie-friendly artistic style abstraction by introducing three core ideas to the traditional CNN based method:

- We propose a neural style architecture that is fully based on the gradient of images. The edge-aware nature of gradient learning provides constraints on edges to eliminating various distortions.
- We introduce color confidence in the reconstruction part to maintain the visual fidelity of the skin color in result images. The reconstruction step exploits both the structure and color information of the original selfies, which ensures the naturalness of the result.
- We initiate the exploration of using perceptual loss directly on gradient domain to enhance the extent of stylization when learning abstraction of diverse style.

3.1. Gradient Constraints and Objectives

One simple strategy to learn the artistic abstraction is that to train the neural network by directly minimizing the summed up pixel-wise loss in RGB channels.

$$\|\mathcal{F}_{\mathbf{W}}(I) - \mathcal{L}\|^2 \tag{1}$$

However, this attempt to maximize objective function directly on the color domain will inevitably lead to the problem of insensitivity to the edge structures. One example of the problems is shown in figure 1, where the shape of the human face in the stylized results significantly diverts from the original selfie images. The skin color is also poorly represented and without smoothness and naturalness. In our further analysis of training

in the color domain and gradient domain, we find that the gradient of stylistic reference images meaningfully diverts from the gradient of the original images, as shown in figure 3.

In order to better evaluate the significance of training in gradient domain, we adopt the biliteral edge-aware filter proposed in Xu et al. (2015). The edge-aware filter smoothes out most of the detailed structure but preserves important edges. Another observation shows that in most of the previous defective samples, the deformation and color shift problem occurs nondeterministically. For example, selfies of the same person taken from slightly different angles can lead to very different facial edges in stylistic patches. In the training process, our method mostly learns those edge-preserved abstractions of the candidate patches in the gradient domain. It is possible to diminish the effect of the defective patches in gradient domain and circumvent the visual drawbacks appearing in previous examples.

With the above understanding, we define the objective function on ∇I rather than I. Considering that most edge-aware operators can produce the same effects even if we rotate the input image by 90 degrees, we use both the vertical gradient $\nabla I_{\mathbf{V}}$ and horizontal gradient $\nabla I_{\mathbf{H}}$ in our training process. Here we denote ∂I as the channel-wise combination of vertical gradient $\nabla I_{\mathbf{V}}$ and horizontal gradient $\nabla I_{\mathbf{H}}$.

Now given D training image pairs $(I_0, \mathcal{L}(I_0))$, $(I_1, \mathcal{L}(I_1))$, $\cdots$, $(I_{D-1}, \mathcal{L}(I_{D-1}))$ of the original selfie images and corresponding unsatisfying stylized reference images, we aim to minimize

$$\frac{1}{D}\sum_{i}\left\{\frac{1}{2}\left\|\mathcal{F}_{\mathbf{W}}(\partial I_i) - \partial\mathcal{L}(I_i)\right\|^2\right\} \tag{2}$$

where $\{\partial I_i, \partial\mathcal{L}(I_i)\}$ denotes the traning example pair in gradient domain. By minimizing the objective function in gradient domain, structural content includes facial edges can be specially emphasized and carefully preserved in the traning process of our framwork.

3.2. Perceptual Loss on Gradient Domain

We make an attempt to apply perceptual loss directly on gradient domain, based on our finding that only using the pixel-wise loss on gradient domain can be restrictive. Perceptual loss enhances the stylization process by extracting high-level semantic representations of edges. Those edges can be important information like brushstrokes in paintings. It is proved that gradient-level semantic information is quite meaningful for our task. An illustration of result with and without the perceptual loss is shown in figure 5.

Following the concept at Johnson et al. (2016), we define the perceptual loss based on VGG-16 image classification network pretrained on ImageNet. The preceptual loss is defined as the l_2-norm between feature representations of the reference image I and the output stylized image $\mathcal{F}_{\mathbf{W}}(I)$.

$$\mathcal{L}_{feat} = \frac{1}{\mathcal{C}_j\mathcal{H}_j\mathcal{W}_j}\left\|\psi_j(I) - \psi_j(\mathcal{F}_{\mathbf{W}}(I))\right\| \tag{3}$$

where $\psi_j()$ denotes the feature map from the j-th VGG-16 convolutional layer and $\mathcal{C}_j$, $\mathcal{H}_j$, $\mathcal{W}_j$ are the number, height and width of the feature maps, respectively. Here we only use the conv3-3 layer for the final output of style representation. We calculate perceptual loss from the euclidean distance between the two outputs of conv3-3 layer.

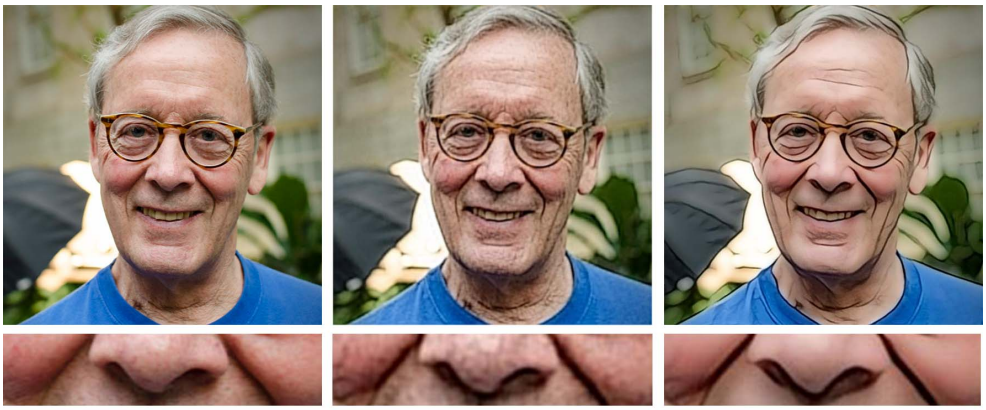

Figure 5: Perceptual loss on gradient domain: (a) The original photorealistic image (b) Training result with only pixel loss (c) Training result with both pixel loss and perceptual loss. Without perceptual loss applied on gradient domain, stylization becomes less abstract and more defective as shown in (b), in which the skin smoothness and stylization quality is not as satisfying as (c).

3.3. Total Loss for Training

We formulate the artistic abstraction learning objective function, combing all two loss components together.

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{pixel} + \beta \mathcal{L}_{feat} \tag{4}$$

where α and β are the corresponding loss weights for pixel loss and style loss.

3.4. Network Architecture

The overall network architecture is illustrated in figure 4. The network is mainly constructed of two parts: the first part takes the gradients of selfie images as input, using continuous convolutional layers with ReLU as feature extraction, then calculate the pixel-wise loss after a sub-pixel upsampling operation. In the second part, the result of the first part will be passed to a VGG-16 network, together with gradients of the reference stylized images. Noted that the weights of the VGG-16 network are fixed in the whole training procedure. After calculating perceptual loss from the output of conv3-3 layer, total loss will be summed up based on the optimized loss weights.

In this network, our architecture also embeds several micro-designs to mitigate checkerboard artifacts. We implement the sub-pixel convolution layer proposed in Shi et al. (2016) instead of deconvolution layer to avoid those artifacts in uneven deconvolution. The sub-pixel upsampling module includes an NNU layer which duplicates the input $N \times N$ times by nearest neighbor strategy, with a masking operation to map the sub-pixels into the corresponding position in the final upsampled feature map.

3.5. Training Details

For the training data, we use a set of selfie images gathered from Flickr as the original images and their corresponding stylistic versions generated from Prisma as the style references. We randomly collect 64×64 patches from selfie images and the corresponding style patches. In

Figure 6: Experiments on people of different age, gender and race: The first column are the original images, and the second column are the stylistic references. The third column are our results. Please zoom in on monitor for better comparison.

our training process, the loss weight α and β are set to 10000 and 10 respectively, to ensure a smooth decrease in both $\mathcal{L}_{pxiel}$ and $\mathcal{L}_{feat}$.

The network was trained on *Nvidia Titan* X GPU for 100K iterations using a batch size of 10. The learning rate was set to be 10^{-8} in first 50K iterations and 10^{-9} in the second 50K iterations. We use Adam algorithm to minimize the total loss shown in eq. (4). The training procedure took about two hours, and the experimental setup was identical in all of the experiments.

Figure 7: Experiments on non-facial scenarios. Note that the training set is totally based on facial images and our method has impressive generalization capability for all cases in which structural realism and color consistency are considered.

3.6. Image Reconstruction

We denote by S our final output gradient map. To maximize the color naturalness and structural realism of human faces in our output, the reconstruction step also exploits the structural information and color content in the input image to guide smoothing in gradient domain. We thus introduce two terms adding together as

$$\|S-I\|+\lambda\left\{\left\|\partial_x S-\mathcal{F}'_{\mathbf{W}}(\partial_x I)\right\|^2+\|\partial_y S-\mathcal{F}_{\mathbf{W}}(\partial_y I)\|^2\right\} \tag{5}$$

where $\|S-I\|$ is the color confidence to use the input image to guide the smoothed image reconstruction. The second term is the common one to use the gradient result both in horizontal and vertical axis. λ is another parameter balancing the two terms.

Note that λ is the balancing factor of color information and structural information in the reconstruction step. This value varies for diverting scenarios of input. Here we perform a greedy search in the result images, which is applied to the testing set of selfie images. In the color recovery process of our experiments, the value of λ is set to be 10.

Figure 8: Our trail of removing checkerboard artifact: (a) The original photorealistic image (b) Training result with deconvolution layer (c) Training result with sub-pixel upsampling module. The checkerboard effects in (b) are eliminated in (c) since we adopt artifact-free architecture in our network setting.

4. Experiments and Applications

We use our framework to learn the selfie-friendly artistic abstraction of a number of real-life stylization effects, including these popular effects generated by Prisma. Although their original algorithms vary in computational cost, our optimized selfie-friendly model can generate stylized images in a universal fast speed, while largely improving the visual attractiveness of the abstraction of facial features. The testing step takes images of size 512×512 as input, and each image takes on average 0.05 second to process using unoptimized MATLAB code.

4.1. Visual Quality Assessment

We conducted two user studies to validate our work. We assessed the results generated by our method, L0 smooth filter Xu et al. (2011), bilateral grid processing Chen et al. (2007), Deep Analogy Liao et al. (2017), CNNMRF Li and Wand (2016), and Fast Neural Style Johnson et al. (2016). To ensure a relatively fair assessment, we used color preservation technique proposed in Gatys et al. (2016a) for all other methods. The survey was conducted on a set of 100 selfie images and user needed to vote for the best method in terms of facial realism and overall preference. Our method won among those methods with more than 50% of the votes. The detail of the result is shown in figure 9.

To ensure our method is robust for selfies in all circumstances (i.e. age, gender, race and etc.), we use selfies of people from diverse backgrounds. As shown in figure 6, our experiment generates convincing results, with the output stylistic image preserving the substantial visual feature of different facial identities, showing that our framework can be adapted to various groups of people, regardless of facial feature differences arouse from their age, gender and race.

4.2. Artifact Removal

At the initial trial of our experiment, we simply use deconvolution layer to upsample the feature map in the first part of our network. The kernel size and stride length we used were also not delicately checked and some of the kernel size used is not dividable by stride in the same convolution layer. The initial result was dramatically affected by checkerboard artifacts, with bothering checkerboard artifacts, as shown in figure 8.

One reason is that overlapping occurs when kernel size is not divisible by stride, which often used as the upsampling factor in deconvolution operation Aitken et al. (2017); Odena et al. (2016). After using the sub-pixel module in the first part of our network and carefully checking the kernel size all convolution layers, these annoying artifacts have been mostly eliminated in results generated from our framework.

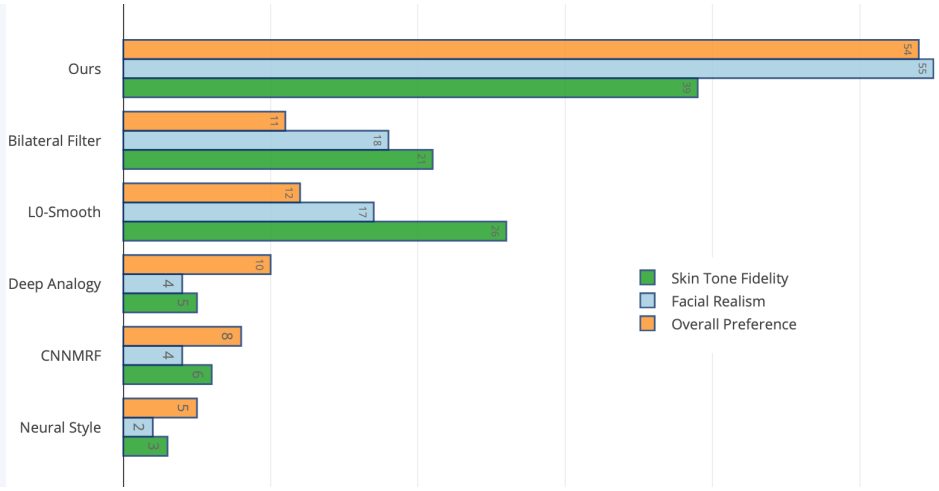

Figure 9: User survey confirming that our algorithm produces faithful results in both structural realism, skin tone fidelity and overall preference assessment.

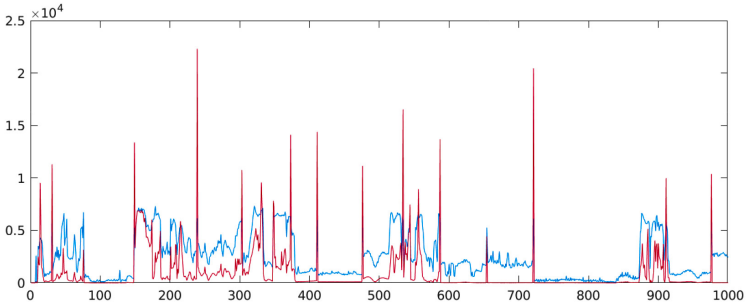

Figure 10: Inter-frame consistency test of our method (red) and the reference method (blue). The y axis is calculated by the MSE of continous frames. Our method demostrates higher consistency in multiple sample video footages. Note that the peaks in our method are related with scene change in the original footage.

4.3. Style Generalization

In this section, we use a large set of images downloaded from Flickr, including landscapes, buildings, and objects for generalization testing. Although the network is trained on selfie dataset, our model still demonstrates impressive generalization capability when dealing with various scenes. The result images embody salient style while preserving the original color and structure fidelity. Example of the results is shown in figure 7.

4.4. Extension to Videos

To test the inter-frame consistency, we apply our model to several video clips of diverse contents, each of them contains over 5000 frames. Figure 10 compares the inter-frame consistency of our method and the most recent method in Ruder et al. (2017). Different from existing methods which have severe effect of flicking when applied to videos, our method shows reliable inter-frame consistency. The stylistic videos generated are smooth and consistent, even for rapid inter-frame transitions.

5. Conclusion

In this paper, we proposed a novel artistic stylization framework that specially optimized for human facial images. The method we proposed exploits the structural information in the gradient domain to preserve structure realism, separates chromatic information from gradient information in the learning process. Our method uses color reconstruction to carefully preserve the skin tone of facial images. The gradient approach explicitly takes advantage of statistical properties in gradient domain to eliminate mismatched patches appearing in the style samples. From our experiments, we showed that our method can generate stylized selfie images qualified as more attractive in both visual and aesthetic assessments.

References

Andrew P. Aitken, Christian Ledig, Lucas Theis, Jose Caballero, Zehan Wang, and Wenzhe Shi. Checkerboard artifact free sub-pixel convolution. 2017.

Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. Coherent online video style transfer. In *ICCV*, 2017.

Hong Chen, Ying-Qing Xu, Heung-Yeung Shum, Song-Chun Zhu, and Nan-Ning Zheng. Example-based facial sketch generation with non-parametric sampling. In *ICCV*, 2001.

Jiawen Chen, Sylvain Paris, and Frédo Durand. Real-time edge-aware image processing with the bilateral grid. SIGGRAPH '07, 2007.

Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer, 2001.

L. A. Gatys, M. Bethge, A. Hertzmann, and E. Shechtman. Preserving color in neural artistic style transfer. Technical report, 2016a.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016b.

Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *CVPR*, 2017.

Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *CVPR*, 2017.

Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, and Mingli Song. Neural style transfer: A review. *CoRR*, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 2012.

Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *CVPR*, 2016.

Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. 2017.

Jingwan Lu, Pedro V. Sander, and Adam Finkelstein. Interactive painterly stylization of images, videos and 3d animations. In *SIGGRAPH*, 2010.

Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *CVPR*, 2017.

Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. Photorealistic style transfer with screened poisson equation. In *BMVC*, 2017.

Meng Meng, Mingtian Zhao, and Song-Chun Zhu. Artistic paper-cut of human portraits. In *Proceedings of the 18th ACM International Conference on Multimedia*, 2010.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.

D. O'Regan and A. C. Kokaram. Skin-aware stylization of video portraits. In *Conference for Visual Media Production*, 2009.

Sylvain Paris, Samuel W. Hasinoff, and Jan Kautz. Local laplacian filters: Edge-aware image processing with a laplacian pyramid. *Commun. ACM*, 58(3):81–91, 2015.

Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos and spherical images. *CoRR*, 2017.

Ahmed Selim, Mohamed Elgharib, and Linda Doyle. Painting style transfer for head portraits using convolutional neural networks. *ACM Trans. Graph.*, 2016.

Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

YiChang Shih, Sylvain Paris, Connelly Barnes, William T. Freeman, and Fredo Durand. Style transfer for headshot portraits. In *Proceedings of SIGGRAPH*, 2014.

Diego A. Socolinsky. Dynamic range constraints in image fusion and visualization. In *Proceedings of Signal and Image Processing*, 2012.

Tanasai Sucontphunt. *3D Artistic Face Transformation with Identity Preservation.* 2014.

Li Xu, Cewu Lu, Yi Xu, and Jiaya Jia. Image smoothing via l0 gradient minimization. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2011.

Li Xu, Jimmy Ren, Qiong Yan, Renjie Liao, and Jiaya Jia. Deep edge-aware filters. In *ICML*, 2015.

Qi Zhang, Xiaoyong Shen, Li Xu, and Jiaya Jia. Rolling guidance filter. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV*, 2014.

Appendix A. Appendix

Figure 11: More experiments on non-facial scenarios. In each set of images, the upper one is the photo-realistic image as input, the bottom one is the stylistic output generated by our method. To know more about the generalization and inter-frame consistency comparison on videos, please visit the online demo at https://youtu.be/0AsY26MHih4.

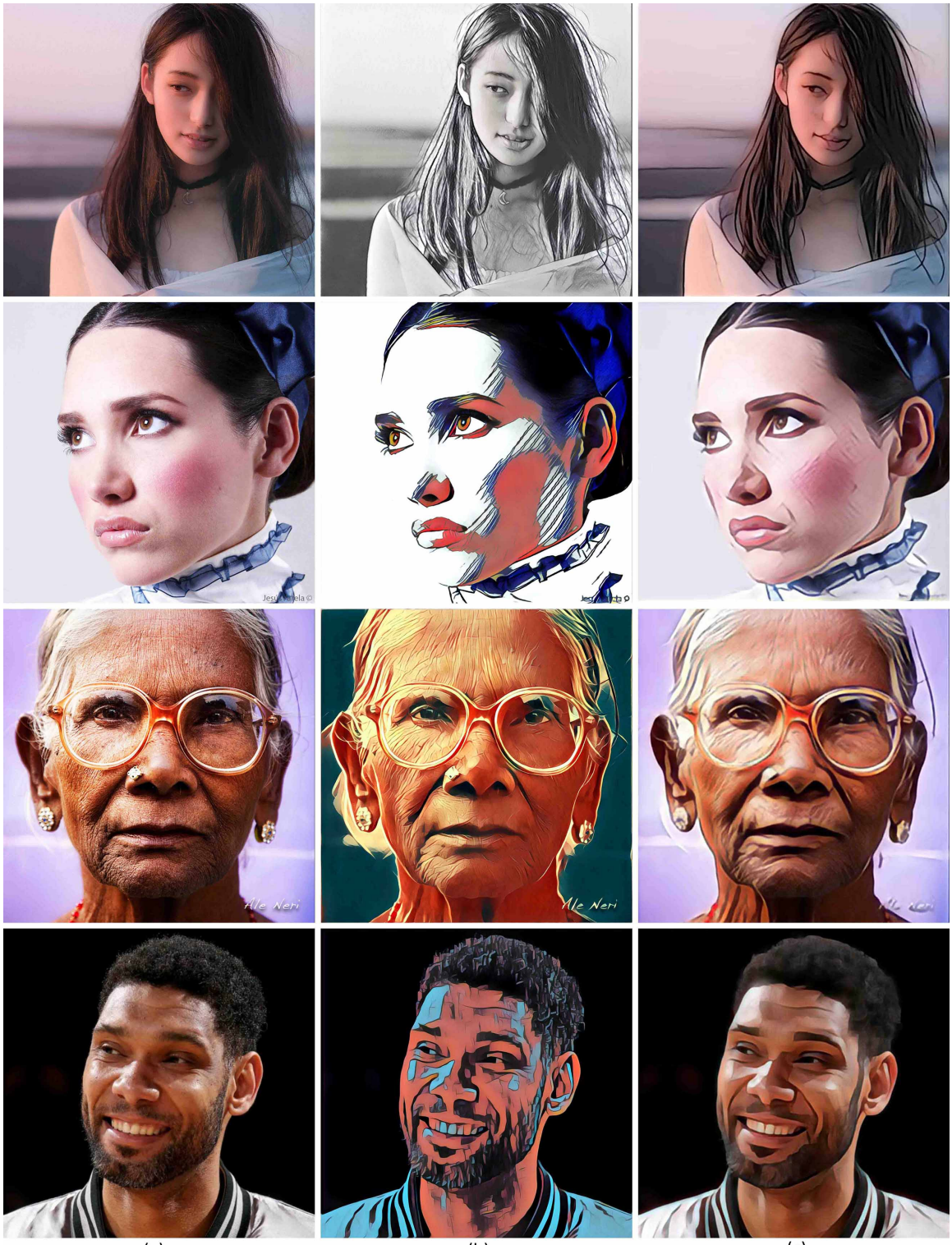

Figure 12: More experiments demonstrating different styles: The first row is result with style 'curly hair' similar to sketches (note that the edge of facial images), the second and the third row are the style 'aviator' with salient edges. The fourth row is another style 'composition', which looks more similar to oil painting.