MULTI-VIEW ARBITRARY STYLE TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we introduce pioneering algorithms for multi-view arbitrary style transfer. Multi-view arbitrary style transfer is an advanced study of the conventional monocular arbitrary style transfer, which aims to preserve the consistent style on the common region across the given arbitrary number of views. We intend to address the multi-view inconsistency problem by minimizing the difference in the color and feature values of the corresponding regions. However, the conventional feature extractors generally produce a feature vector of a point from its rectangular local patch, and such local patches are misaligned across the views in a multi-view environment due to various camera poses. Thus, even if we assimilate the feature vectors of the corresponding pixels, since the spatial distribution of the surrounding feature vectors within their local patches are different and decoding such misaligned patches induces misaligned brushstrokes and geometric patterns, these feature vectors can be decoded to distinctive style texture. Based on the observation, we intend to interpret this challenging problem in terms of the photometric inconsistency and the stroke inconsistency. We propose a photometric consistency loss, which directly enforces the geometrically consistent style texture across the view, and a stroke consistency loss, which matches the characteristics and directions of the brushstrokes by aligning the local patches of the corresponding pixels before minimizing feature deviation. Then, We construct an optimization-based multi-view arbitrary style transfer framework (MVAST-O) with photometric and stroke consistency losses and extend it to a feed-forward framework (MVAST-FF) to overcome the chronic computational inefficiency issue of optimization-based algorithms. We validate our methods on the DTU dataset, a large-scale multi-view stereo dataset, and confirmed the superiority on preserving appearance consistency throughout the stylized multi-view images.

1 INTRODUCTION

Multi-view vision has recently attracted new attention with the development of computing devices and deep learning skills. Several works (Yao et al., 2018; Chen et al., 2019) demonstrate promising performances in multi-view stereo. In industry, autonomous mobilities and portable devices have adopted multi-camera systems for safety and satisfactory performances. Current trend of computer vision technology is to better understand the 3D world through multi-view systems, and thus the conventional monocular or stereoscopic vision tasks are being extended to fit the multi-view environment. Style transfer is also one of such tasks that needs to be extended accordingly

Style transfer is an important image manipulation task aiming at synthesizing an image that preserves the semantics of the content image while assimilating the texture of the input image to the style image. Starting from the pioneering optimization-based approach of Gatys et al. (2015; 2016), subsequent works attempt to improve the computation and memory efficiency as well as the stylization quality (Johnson et al., 2016), and generalize the algorithms even to the unseen style images (Li et al., 2017b; Huang & Belongie, 2017; Sheng et al., 2018). However, the possibility of applying neural style transfer to multi-view images has not yet been sufficiently explored.

As a pioneer work to the novel task, we introduce algorithms for multi-view arbitrary style transfer: an optimization-based approach and a feed-forward approach. The core challenge of the task is to preserve the consistency of the appearance of the corresponding regions across the views while stylizing the multi-view scenes with arbitrary style images. Although there have been many studies in field of monocular style transfer, there are many problems in applying these methods to multi-

Figure 1: Visualization of the inconsistent stylized texture in spatially corresponding areas. (a) Given 3-view images and a style image, we compared the stylization results of conventional monocular style transfer method (first row) and our method (second row).

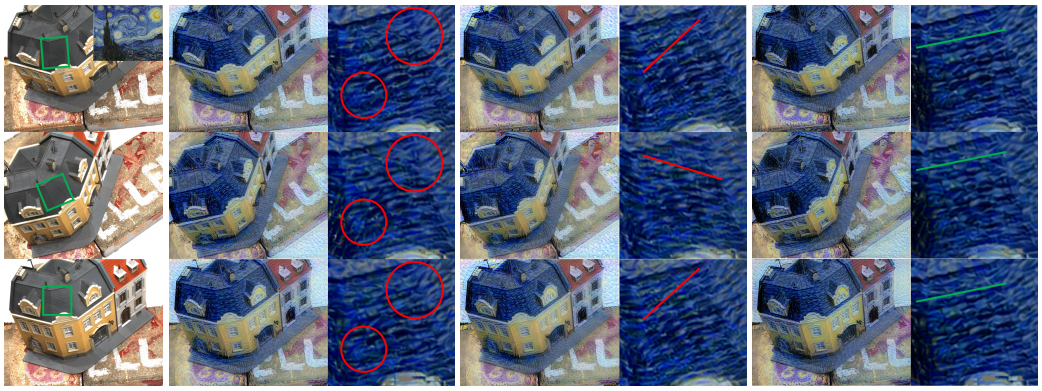
view style transfer. As shown in Fig.1, the conventional monocular style transfer method often fails to maintain consistency of appearance across views. First of all, we introduce a photometric consistency loss, which directly minimizes the deviation of the stylized appearance between the corresponding pixels. However, since the photometric consistency loss only constraints the stylization results pixel-wise regardless of the local geometries, the stylized multi-view scenes often show local geometric inconsistencies in the common areas due to feature differences. Such deviations in common areas should also be minimized.

However, we observed that directly minimizing the feature inconsistency across the views rather cause the misalignment of the transferred brushstroke. As shown in Fig. 2, despite the feature deviation reduction, the corresponding strokes of the stylized scenes are often misaligned and oriented in inconsistent directions, which we call stroke inconsistency. Such stroke inconsistency originates from the projective distortions between scenes and the characteristic of the conventional neural networks, which can only extract features from rectangular patches. Specifically, since the rectangular patches surrounding the corresponding points cannot avoid rotational misalignment across views and the characteristics of the stylized strokes are highly related from the feature values, the feature map of a given view cannot properly constrain the stroke consistency on other views' stylized images. Thus, we introduce a stroke consistency loss, which focuses on generating the modified feature maps extracted from aligned patches and minimizes the differences between feature maps of the corresponding parts or areas across views.

With the proposed losses, we first construct an optimization-based multi-view arbitrary style transfer framework (MVAST-O), adopting an optimization-based stylization method as a backbone algorithm. Moreover, we also extend MVAST-O to a feed-forward multi-view arbitrary style transfer framework (MVAST-FF) to resolve the chronic computational inefficiency issue of optimization-based algorithms. We extensively conducted quantitative comparisons and analyses in Section 4 and appendix, and confirmed that our methods show significantly better performances in preserving geometrically consistent stylized textures on stylized multi-view scenes.

To summarize, our contributions are fourfold:

- We explore a novel multi-view arbitrary style transfer task. We additionally claim a new issue called stroke inconsistency caused by the fundamental setting of multi-view systems.
- We propose an optimization-based multi-view arbitrary style transfer (MVAST-O) regarding the photometric consistency and stroke consistency losses. Here, the stroke consistency loss aims to align the directions of the transferred brushstrokes across the views.
- To overcome the chronic computational inefficiency of optimization-based algorithms, we extend the core ideas to a feed-forward approach and construct a feed-forward multi-view arbitrary style transfer network (MVAST-FF).
- We extensively conducted experiments and analyses with the multi-view dataset and confirmed the qualitative superiority of our methods.



(a) Input images (b) Photometric only (c) Photometric + Feature (d) Photometric + Stroke

Figure 2: Qualitative comparisons of constraint effects between various consistency losses.

2 Related Work

Multi-view style transfer Few works (Chen et al., 2018; Gong et al., 2018) have recently explored stereoscopic style transfer, attempting to preserve geometric consistency between stereoscopic pairs by aggregating each view features to the other view via a disparity based warp function. However, these approaches are limited to two scenes only and hard to adopt for multi-camera or multi-lens devices. Unlike the previous work, we focus on expanding the applicability of style transfer to an arbitrary number of views.

Multi-view vision Multi-view vision is a promising topic that bridges the 2D images to the 3D world. Multi-view stereo (MVS) has been researched for a long time, and recently learning-based MVS networks (Yao et al., 2018; Chen et al., 2019) demonstrate excellent performances on large-scale MVS benchmarks (Jensen et al., 2014). With advances of the MVS methods, Nam et al. (2019) attempts to capture high fidelity hair geometry at strand-level using the estimated point cloud. There are also some works that focus on surface refinement, 3D edge reconstruction, and text co-detection in the multi-view environment. However, few works have studied multi-view style transfer, and most of them studied only for the stereoscopic case (Chen et al., 2018; Gong et al., 2018). In this paper, we focus on stylizing an arbitrary number of views without visual fatigue.

3 Method

3.1 Optimization-based multi-view arbitrary style transfer (MVAST-O)

Fundamental stylization losses Though we are focusing on preserving the multi-view consistency, multi-view arbitrary style transfer algorithm should meet the necessity conditions of style transfer. Given N-view images $\{I_v\}_{v=1}^{N}$, we utilize the standard style reconstruction loss L_s^v and content reconstruction loss L_c^v for any view index $v \in \{1, ...N\}$:

$$\begin{aligned} L_c^v(O_v, I_v) &= \sum_{l \in l_c} \|F^l(O_v) - F^l(I_v)\|^2, \\ L_s^v(O_v, I_v) &= \sum_{l \in l_s} \|G(F^l(O_v)) - G(F^l(I_v)\|^2, \end{aligned} \tag{1}$$

where O_v denotes the stylized result of the v-th scene, $F^l(\cdot)$ is the feature maps function for the layer l of the feature extractor and $G(\cdot)$ is the gram matrix function. In addition, we apply the total variation loss to reduce the rough texture:

$$L_{tv}^v(O_v) = \|\nabla_x O_v\| + \|\nabla_y O_v\|. \tag{2}$$

Occlusion estimation Since considerable occlusions occur in a multi-view setting and significantly affect constraining the multi-view consistency, we introduce an occlusion estimation method

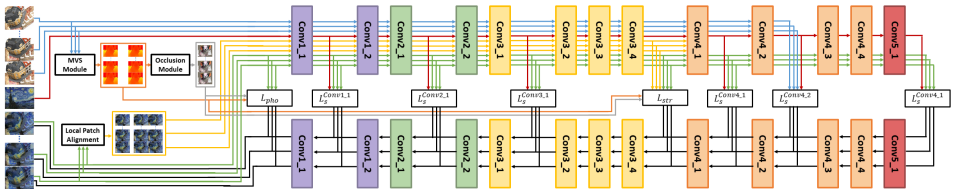


Figure 3: An overall framework of MVAST-O. Our framework consists of a VGG-16 encoder, an occlusion estimation module, an MVS module, and a Local Patch Alignment (LPA) module. The stylized results are iteratively optimized through the back-propagation of the total loss.

to filter occluded pixels. The key idea is to measure the closest distance between the unprojected pixels of each view in the world frame and check if the distance exceeds a certain threshold. This approach makes sense since if a occluded pixel $p \in A(I_v)$ does not have correspondence in $A(I_{v'})$ for a given view pair (v, v'), Chamfer distance between p and the unprojected point cloud of $I_{v'}$ must be large. In practice, given multi-view scenes $\{I_v\}_{v=1}^N$, depth maps $\{d_v\}_{v=1}^N$, and camera parameters $\{K_v, R_v, t_v\}_{v=1}^N$, we estimate an occlusion mask $M_{v,v}$ of I_v with respect to $I_{v'}$ by:

$$M_{v,v'}^{\epsilon}(p) = \begin{cases} 0 & \text{if } \min\limits_{q \in A(I_{v'})} \|W_{v \to world}(p) - W_{v' \to world}(q)\| < \epsilon, \\ 1 & \text{otherwise,} \end{cases}$$
$$W_{v \to world}(p) = R_v^{-1}(d_v(p) K_v^{-1} p - t_v), \tag{3}$$

for all $p \in A(I_v)$ and $1 \leq v, v' \leq N$, where $W_{v \to world}(\cdot)$ is a warp function from the v-th camera frame to the world frame, $A(\cdot)$ is a set of visible pixels on the image plane, and ϵ is a threshold. We visualized the qualitative results of the occlusion mask in Fig. **??** and analyzed on Section 4.3.

Photometric consistency loss As the fundamental stylization losses optimize, each scene individually assimilates the texture into the style image regardless of its multi-view correspondences, and eventually the corresponding regions gradually diversify their appearances. We directly minimize the deviation of corresponding pixel values in the stylized results to handle the inconsistency, which we call a photometric consistency loss. Suppose we have stylization results $\{O_v\}_{v=1}^N$ of each view, then the photometric consistency between O_v and $O_{v'}$ can be obtained by:

$$L_{pho}^{v,v'}(O_v, O_{v'}) = \sum_{p \in A(O_v)} M_{v,v'}^{\epsilon}(p) \|O_v(p) - O_{v'}(W_{v \to v'}(p))\|^2, \tag{4}$$

for all $1 \leq v, v' \leq N$ where $W_{v \to v'}(\cdot)$ is a warp function from the v-th camera frame to v'-th camera frame, which can be derived by the following projective equality:

$$W_{v \to v'}(p) \sim K_{v'}(R_{v'} W_{v \to world}(p) + t_{v'}), \tag{5}$$

for all $p \in A(O_v)$, where '$\sim$' denotes the projective equality.

Stroke consistency loss We empirically observed that directly minimizing the differences between stylized view scenes or their feature maps cannot sufficiently align the direction or assimilate the style of the corresponding brushstrokes, or even wrongly align the brushstroke due to the different camera poses. The main reason is that the style of the transferred brushstroke is determined by its own and the surrounding features, and these features are extracted from the local patches, which are usually misaligned in multi-view setting. Specifically, although a pair of views share common regions, local patches of these regions cannot be aligned due to the the different principle axes of views and the rectangular kernel of the convolution neural networks. Thus, to correctly compare the strokes on the corresponding region, we should align local patches of view images $\{O_v\}_{v=1}^N$ before encoding. To obtain the aligned local patch of $O_{v'}$ with respect to vth camera frame, we warp $O_{v'}$ and complete the occluded region with O_v and $M_{v,v'}^{\epsilon}(p)$ via interpolation:

$$O_{v' \to v}^*(q) = (1 - M_{v,v'}^{\epsilon}(q)) * O_{v'}(W_{v' \to v}(q)) + M_{v,v'}^{\epsilon}(q) O_v(q), \tag{6}$$

for all $q \in A(O_{v'})$ and $1 \leq v, v' \leq N$. The completion process is necessary since the zero values on the occluded regions can cause critical feature corruption during feature extraction. Then, the

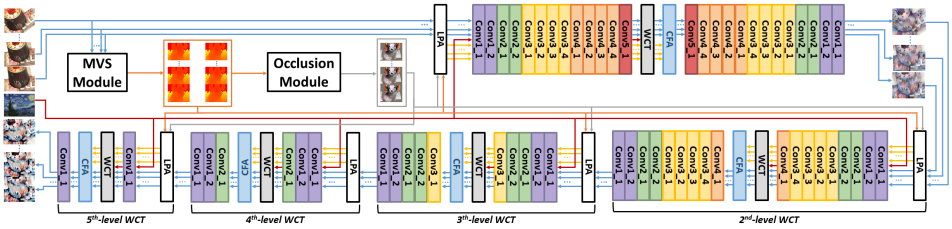


Figure 4: An overall framework of MVAST-FF. Our framework stylizes multi-view scenes in five levels in a coarse-to-fine way. At each level, it aligns local patches through the Local Patch Alignment (LPA) module and applies Camera Frame-wise Average(CFA) to fuse the transformed features.

stroke consistency loss can constrain the inconsistent stroke by minimizing the feature map between O_v and the aligned image $O_{v \to v'}$ for any view pair (v, v'):

$$L_{str}^{v,v'}(O_v, O_{v'}) = \| |F^i(O_v) - F^i(O^*_{v' \to v})\|^2. \tag{7}$$

Multi-view stereo (MVS) module Depth maps are essential for the photometric and stroke consistency losses derivation, as shown in the previous formulas. However, depth sensors like LiDAR sensors are not easily used in real life due to their cost-inefficiency and large size, so accurate depth may not always be available. Thus, to alleviate the dependence on ground-truth depth map, we utilize the multi-view stereo module to estimate the depth maps from given images and camera parameters. Our module consists of a feed-forward multi-view stereo network and the post-processing, and we use the default MVSNet (Yao et al., 2018). However, directly using the predicted depth maps causes incorrect correspondence across the views due to the prediction error, and the photometric and stroke consistency losses will assimilate the stylized texture of these wrongly related regions, causing fatal stylization defects. Inspired by Galliani et al. (2015), we filter out pixels with inaccurate depth by warping each pixel to its corresponding pixels of other views using the predicted depth and compare the depth values.

Overall framework and optimization problem Figure 3 describes the overall framework of MVAST-O. It employs VGG-16 Simonyan & Zisserman (2015) as an encoder and iteratively reconstructs the stylized output with style loss, content loss, tv loss, and two proposed multi-view consistency losses. We can estimates reliable depth maps with raw images from the MVS module, or we directly utilize ground-truth depth maps for unprojection. We compute all the losses for each view and sum with proper loss weights:

$$\begin{aligned} L_{total}\left(\{O_v, I_v\}_{v=1}^N\right) = \sum_{v=1}^{N} \Big(& \alpha L_c^v(O_v, I_v) + \beta L_s^v(O_v, I_v) + \gamma L_{tv}^v(O_v) \\ & + \sum_{v'=1, v' \neq v}^{N} \left(\delta L_{pho}^{v,v'}(O_v, O_{v'}) + \zeta L_{str}^{v,v'}(O_v, O_{v'}) \right) \Big) \end{aligned} \tag{8}$$

Finally, we iteratively update the stylized results $\{O_v\}_{v=1}^N$ through total loss minimization.

3.2 Feed-forward based multi-view arbitrary style transfer (MVAST-FF)

Regardless of the rich stylization performance of MVAST-O, optimization-based algorithms require high computational costs, hindering real-time applications. Therefore, it is necessary to approximate the optimization-based algorithm in a feed-forward manner. While MVAST-O can directly update the stylized results $\{O_v\}_{v=1}^N$ iteratively, the feed-forward multi-view style transfer network should manipulate all the inconsistently transferred stylized texture at one pass. However, the various camera poses induce a trade-off between stylization quality and multi-view consistency with respect to the level of feature transformation in the multi-view setting. In detail, we can expect a global but coarse stylizing effect from deep feature transformations. However, even if we assimilate the corresponding feature vectors of the common non-occluded region, the spatial distribution of the surrounding feature vectors within the local patches of each corresponding region cannot be the

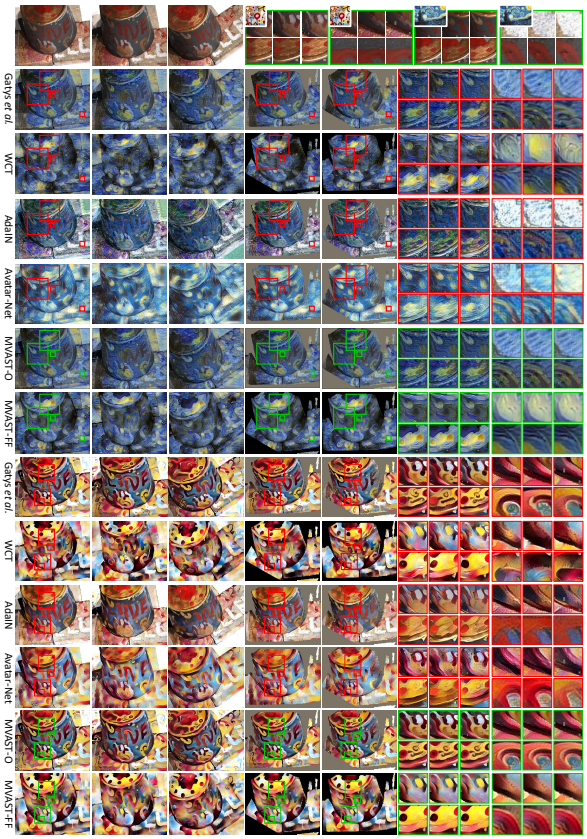


Figure 5: Qualitative comparisons with arbitrary style transfer baselines and our methods.

same due to differences in camera poses. These gaps are magnified during the deep decoding process, causing significant appearance inconsistency. On the other hand, assimilating shallow features can guarantee consistent textures since the shallow decoding process cannot significantly amplify the inconsistency, though only fine but local stylization effects can be expected.

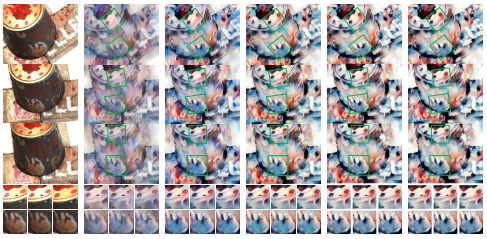
We introduce a feed-forward framework for multi-view arbitrary style transfer, which reflects the core ideas of photometric and stroke consistency losses for a multi-view feature fusion process while maintaining adequate stylization quality. To achieve satisfactory effects on both sides, we use a coarse-to-fine stylization network for macro style transfer and utilize the multi-view feature fusion for progressive inconsistency reduction. In practice, we adopt multi-level whitening and colorization transformation (ML-WCT) as the backbone coarse-to-fine stylization network. For each coarse-to-fine level, we align local patches before encoding, match the second-order statistics of encoded $\{I_v\}_{v=1}^N$ and $\{I^*_{v\to v'}\}_{v=1}^N$ with the style feature, and average them with respect to each camera frame before decoding. The overall framework is described in Fig. 4.

4 Experiments

We additionally reported experimental results and analyses in the appendix due to the limited space. Moreover, to help compare the multi-view consistency of the stylized scenes, we provide GIF files of the figures from the perspective of a fixed camera poses in the supplementary material.

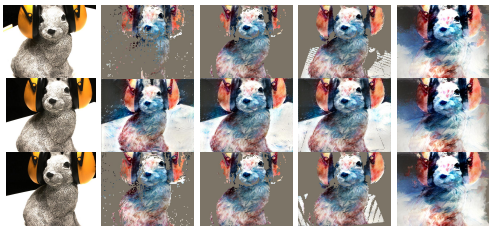
4.1 Experimental Setup

Datasets Since there is no standard benchmark for multi-view arbitrary style transfer, we employed multi-view images with camera parameters from the multi-view stereo dataset. DTU (Aanæs



(a) (b) (c) (d) (e) (f)

Figure 6: Coarse-to-fine stylization results of MVAST-FF.



(a) (b) (c) (d) (e)

Figure 7: Occlusion maps and stylization results with MVSNet predictions without occlusion map.

et al., 2016) is a large-scale multi-view stereo dataset containing 124 scenes with seven lighting conditions. We used multi-view samples from the DTU dataset for comparisons.

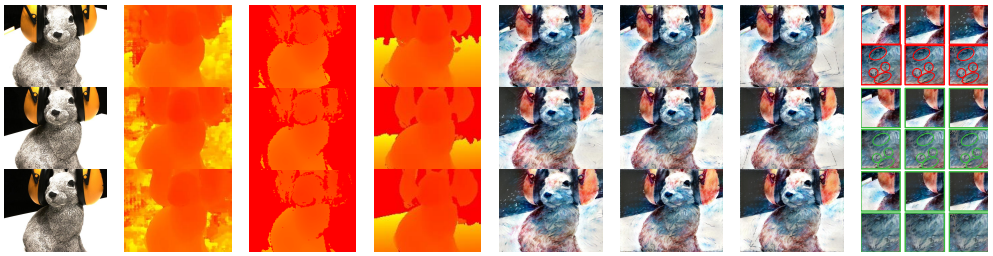
Implementation details We used VGG-16 for the encoder of MVAST-O and the auto-encoder of MVAST-FF, and adopt MVSNet for the initial depth estimation network in the MVS module. Here, VGG-16 is pretrained on ImageNet, and MVSNet is trained on a training split of the DTU dataset according to the training curriculum in the original paper. For the MVAST-O implementation, We applied style reconstruction loss on layer 'conv1_1', 'conv2_1', 'conv3_1', 'conv4_1', 'conv5_1', content reconstruction loss on 'conv4_2', and stroke consistency loss on layer 'conv3_4'. We set 5 for the content reconstruction loss weight, 2×10^3 for each style reconstruction losses, 1×10^{-3} for the total variation loss, 1×10^3 for the photometric consistency loss, and 1×10^1 for the stroke consistency loss. We iteratively updated the outputs by the L-BFGS optimizer (Liu & Nocedal, 1989) with a learning rate 1 for 1000 iterations. For the MVAST-FF implementation, we trained the multi-view WCT (Li et al., 2017a) according to the original paper and followed the MVAST-O settings for all shared modules. All experiments were performed on a single GeFore Titan XP GPU.

4.2 Qualitative Comparisons

We compared our methods with various arbitrary style transfer baselines including Gatys et al. (2016), WCT (Li et al., 2017a), AdaIN (Huang & Belongie, 2017), Avatar-Net (Sheng et al., 2018) to validate the effectiveness on preserving consistent stylized textures across the views. We employed 3-view scenes from *Scan1* split of the DTU dataset for content images, and *Starry Night* and *Candy* for style images. Here, we used ground-truth depth maps to precisely validate the effectiveness of the photometric and stroke consistency regularization. The 1^{st} through 3^{rd} columns of Fig. 5 show stylized results of the three scenes. Though baselines have distinctive stylization properties each other, they seem to change the scenes similarly at first glance. However, as shown in 4^{th} and 5^{th} columns, if we compare the stylized multi-view scenes from a certain fixed camera frame, we can observe several inconsistent regions from the baseline results. For example, as shown in the local patches of 6^{th} and 7^{th} columns, all the baselines show inconsistent color or geometric patterns while our MVAST-O and MVAST-FF preserve the stylized textures and the characteristics and direction of the transferred brushstroke. Specifically, since Gatys et al. and Avatar-Net used to reflects local geometries of the style image, as shown in the 6^{th} column, unique patterns of *candy* were well-described in the stylized results. However, despite being a corresponding region, uncontrolled encoding and decoding processes cause similar but subtly different patterns. Moreover, as shown in the 7^{th} column, while Gatys et al., WCT, and AdaIn distinctively distort or blur and Avatar-Net inconsistently colorize the unique brushstrokes of *starry night*, MVAST-O and MVAST-FF consistently decode the transferred brushstrokes. In additional, we visualized the stylization progress of MVAST-FF by level in Fig 6.

4.3 Analysis

Analysis on multi-view stereo (MVS) module Since the MVS module consists of MVSNet for initial depth prediction and the subsequent post-processing, we qualitatively compare their depth maps in Fig. 8(b) to (d). While MVSNet predicted confidently on the object region, the degenerative



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 8: Qualitative ablation results on MVS module and the induced stylization results.

error occurs in the background region due to the patternless background, which can fatally affect the following multi-view stylization process. On the other hand, filtering out outliers through the consistency check successfully removed wrong predictions, as shown in Fig. 8(c).

Analysis on occlusion estimation From Fig. 7(b) to Fig. 7(d) visualizes the occlusion estimation results derived from MVSNet predictions, MVS module results, and ground-truth depth maps. As expected, pixels in the background region barely found its correspondences due to the unstable depth values of the MVSNet predictions. Thus, most of the background pixels are occluded except some fortunate spots. But, in fact, these few pixels are also errors induced by the accidental correspondences with wrong pixels in other views. In contrast, since our MVS module filtered most of he inconsistent depth prediction values, we can obtain stable occlusion masks. Even compared with ground-truth depth map based occlusion masks, they maintain most of the object regions, which facilitate the subsequent multi-view stylization process.

Analysis on how inaccurate depth map prediction affects consistent multi-view stylization Despite filtering out some pixels with incorrectly predicted depth values, we observed that remaining inaccurate depth predictions can distort the multi-view stylization results. From Fig. 8(e) to Fig. 8(g) show MVAST-O prediction results with MVSNet predictions, MVS module results, and ground-truth maps. Then, we additionally zoomed in on local patches in Fig. 8(h) to clearly visualize the inconsistencies. First of all, as shown in the upper patches of Fig. 8(h), the spotty wrong correspondences in the background cause spatially inconsistent spotty texture. We also observed spatial distortions of the stylized textures on the object from the bottom local patches. This can happen since even though there is an error in the predicted depth value of a pixel, if a pixel near the corresponding pixel in another scene has an incorrect but appropriate depth value so the two pixels can meet closely in the world frame, they can avoid occlusion filtering and be stylized consistently in subsequent processes. In contrast, though stylization results with MVS module also have slight spatial distortion, they show more consistent appearances.

Analysis on effect of estimated occlusion masks To verify the effectiveness of the estimated occlusion mask, we visualized the stylization results of MVAST-O using MVSNet predictions (Fig. 8(b)) without occlusion. Since the occluded pixels are not filtered and thus constrained by consistency losses, the ghost artifacts appeared near the headset, as shown in Fig. 7(e).

5 Conclusion

We explored a novel multi-view arbitrary style transfer task, which has not sufficiently been explored yet. Unlike monocular and stereoscopic style transfer, We observed that minimizing the feature deviation on the common non-occluded region can cause misaligned brushstrokes in a multi-view system. Thus, we claimed the stroke inconsistency problem in the multi-view style transfer. To solve the multi-view arbitrary style transfer task, inspired by the claim, we attempted to minimize the photometric and stroke inconsistencies. We employed these key ideas in an optimization-based approach in the form of the loss functions and a feed-forward approach in the way of the feature fusion process. We demonstrated the successful multi-view consistency preservation through extensive experiments and analyses.

REFERENCES

Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pp. 1–16, 2016.

Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stereoscopic neural style transfer. *CVPR 2018*, 2018.

Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 873–881, 2015.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015. URL http://arxiv.org/abs/1508.06576.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Xinyu Gong, Haozhi Huang, Lin Ma, Fumin Shen, Wei Liu, and Tong Zhang. Neural stereoscopic image style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision(ICCV)*, 2017.

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 406–413. IEEE, 2014.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, 2017a.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 386–396. 2017b.

Dong C. Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Math. Program.*, 45(1-3):503–528, 1989. URL http://dblp.uni-trier.de/db/journals/mp/mp45.html#LiuN89.

Giljoo Nam, Chenglei Wu, Min H. Kim, and Yaser Sheikh. Strand-accurate multi-view hair capture. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*, pp. 1–9, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*, 2018.

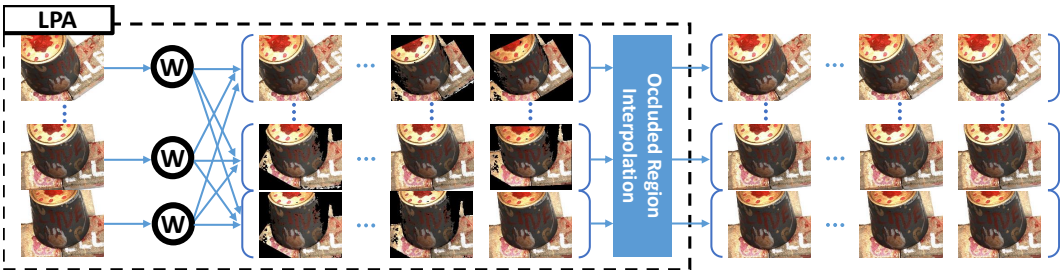


Figure 9: Internal computation descriptions of the Local Patch Alignment (LPA) module essential for stroke consistency regularization.

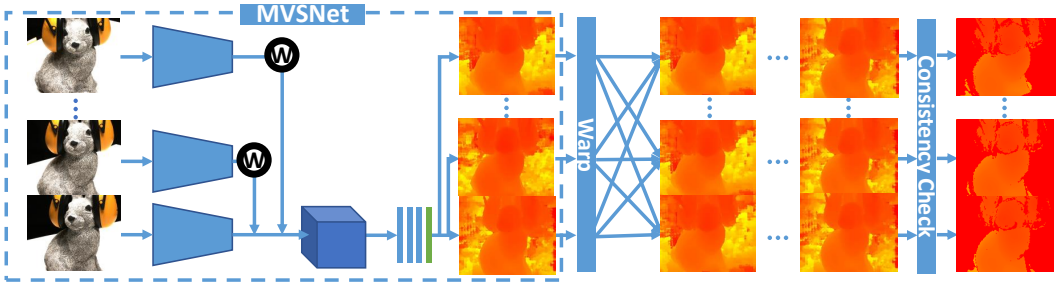


Figure 10: Internal computation descriptions of the Multi-view stereo (MVS) module.

A Appendix

A.1 Additional implementation details

Our visual discriptions of MVAST-O and MVAST-FF in the main script are quit simplified due to its complex structure and the limited space. Thus, we additionally described the detailed computation processof the essential modules, a Local Patch Alignment (LPA) module and a Multi-view Stereo (MVS) module, in Fig. 9 and Fig. 10.

A.2 Additional qualitative results

Additional multi-view stylization results of MVAST-FF Since we visualized limited stylization results with a few multi-view samples and style images, we provide additional results with various style images. Figure 11 and 12 show the qualitative results of MVAST-FF with multi-view sample from *Scan75* and *Scan118* splits of the DTU dataset, respecitvely.

Qualitative comparisons with baselines though animated GIF data It is not easy to find multi-view inconsistencies only with static images since someone needs to simultaneously grasp the appearance deviations across stylized multi-view scenes. Thus, we provides animated GIF data of the Fig. 5 and 8 to better figure out the inconsistent regions.

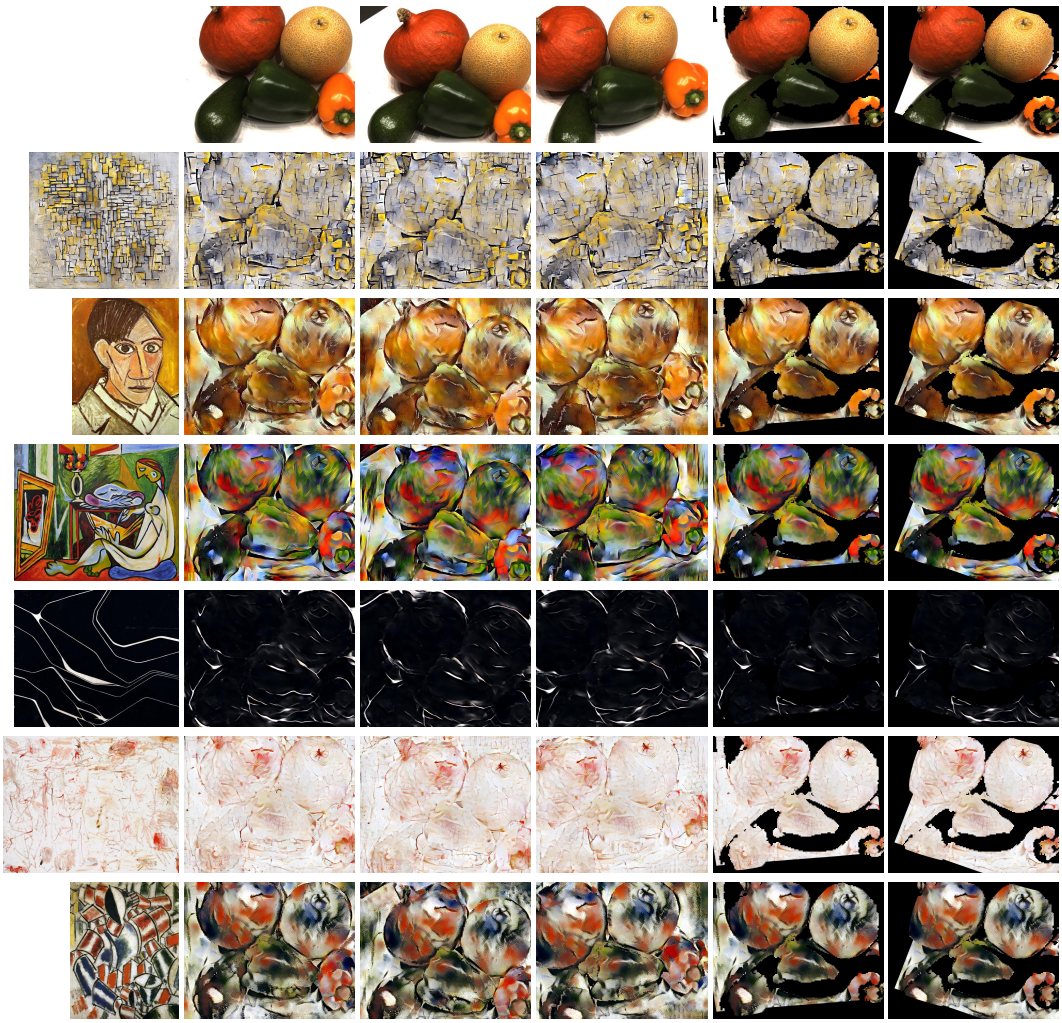


Figure 11: Qualitative results of the feed-forward based multi-view aribtrar style transfer algorithm (MVAST-FF). We used a multi-view scenes from *Scan75* split of the DTU dataset for content images. 4^{th} and 5^{th} columns visualize the second and third scenes in the perspective of camera frame of the first scene.

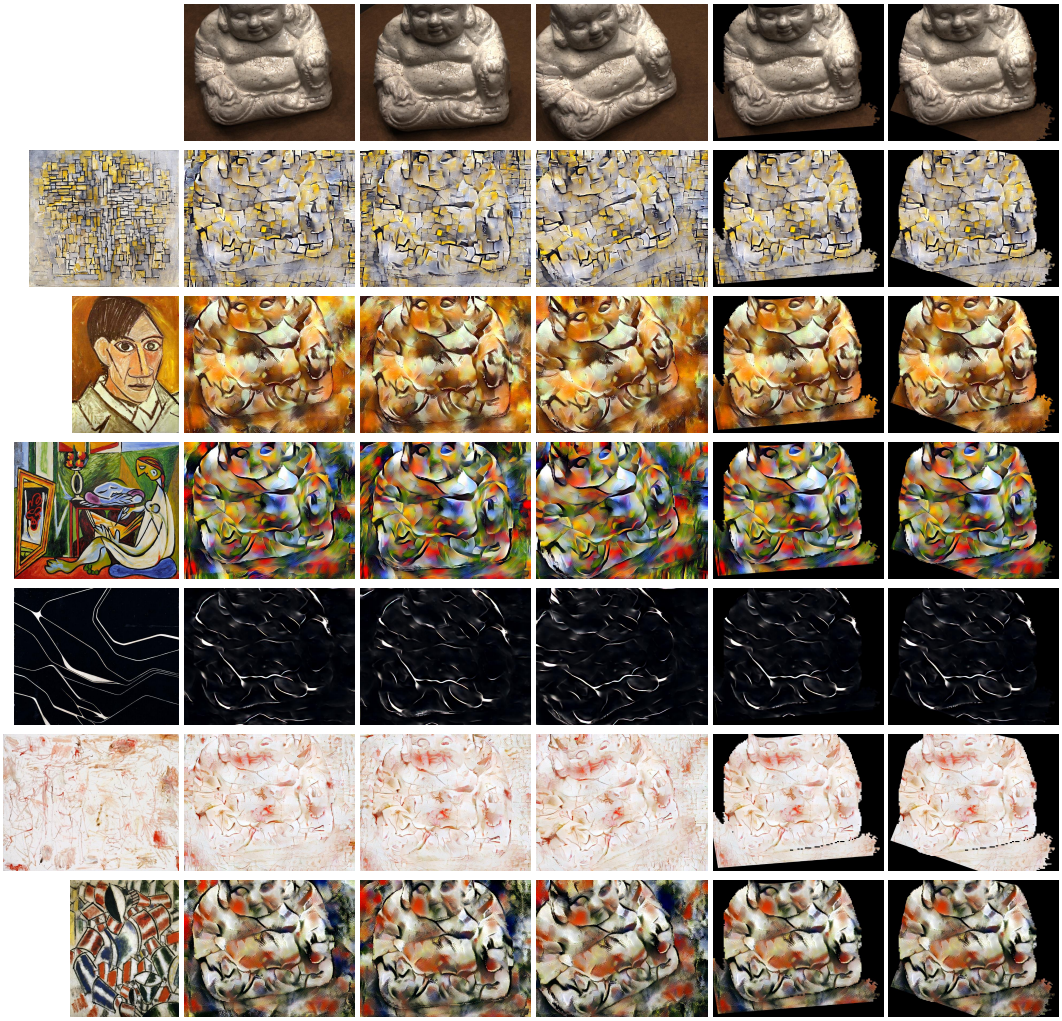


Figure 12: Qualitative results of the feed-forward based multi-view aribtrar style transfer algorithm (MVAST-FF). We used a multi-view scenes from *Scan118* split of the DTU dataset for content images. 4^{th} and 5^{th} columns visualize the second and third scenes in the perspective of camera frame of the first scene.