第 27 回日本バーチャルリアリティ学会大会論文集（2022 年 9 月）

A Method to Style Transfer for Free-Viewpoint Video Generation

深層学習を用いた自由視点映像のスタイル変換

Zhizheng Xiang [1)], Hidehiko Shishido [2)], Itaru Kitahara [2)]

1) Doctoral Program in Empowerment Informatics, University of Tsukuba
(〒305-8573 Tennodai 1-1-1, Tsukuba, Ibaraki, xiang.zhizheng@image.iit.tsukuba.ac.jp)
2) Center for Computational Sciences, University of Tsukuba
(〒305-8577 Tennodai 1-1-1, Tsukuba, Ibaraki,, {shishido | kitahara}@ccs.tsukuba.ac.jp)

Abstract: This paper proposes a straightforward framework that combines 3D free-viewpoint generation based on NeRF (Neural Radiance Fields) with 2D image style transfer. NeRF allows for state-of-the-art results for synthesizing novel views of objects with complicated geometry and appearance. Image style transfer, a deep-learning-based algorithm that separates and recombines content and style of arbitrary images, can generate entertaining artistic 2D images. The proposed system uses pre-trained neural networks to extract content and style information of images and NeRF to synthesize novel stylized views of arbitrary given objects. Moreover, we focus on tackling the inconsistency issue caused by image style transfer since it originally aims to stylize a single image without any view consistency.

Keywords: Style Transfer, Novel View Synthesis, Free-Viewpoint Video, NeRF (Neural Radiance Fields)

1. Introduction

Generating entertaining artwork with the neural network has achieved amazing results in 2D image generation. The goal of style transfer, a branch of image generation, is to capture the style features of a given image and then transfer it to another image while maintaining the semantic content. Although the previous methods perform extremely well in 2D space (image), they usually suffer from a view-inconsistent problem when it comes to 3D space (3D reconstruction based on multi-view images). Hence, to bring style transfer into 3D space and make it shed light, it is important to generate a set of consistent visual images.

As shown in figure 1, our proposed approach can extract the style information of a given image and apply it to a complex real-world 3D scene. In addition, it is also able to synthesize view-consistent stylized images from novel views based on the neural radiance field.

We use pre-trained VGG (Visual Geometry Group) networks [1] as a style features extractor which proved to be an efficient way of performing style transfer [2]. However, the Gram matrix-based style transfer method often generates notable artifacts and suffers from view inconsistent problems when applying it to images taken from different viewpoints. This limitation encourages us to exploit a novel loss function that is better suited for 3D scene style transfer. To be more specific, we add Gaussian noise to the target image to improve the model's robustness toward multi-view images. It demonstrates to be a simple while efficient way to keep the consistency of the generated video after stylizing.

2. Related Work

Feature extractor-based style transfer. Gatys et al. [2] capture the style features by using a pre-trained convolutional neural network (VGG-19 [1]) and compute the loss (a content loss captured by the internal CNN (Convolutional Neural Networks) layer and the style loss represented by the Gram matrix) between style and content images to generate entertaining artistic painting. While the method produces high-quality results, it is extremely time-consuming since each iteration of optimization requires a forward and backward propagation towards generated, content and style images. To address the computational expensive issue, Johnson et al. [3] train a feed-forward convolutional neural network to quickly approximate solutions to the style transfer problem. Our approach is inspired by this method and adds an additional loss item to train the network gaining the ability to keep view consistency.

Generative model-based style transfer. Instead of performing iterative optimization, Generative Adversarial Networks (GAN) [4][5] train a generative model that captures the data distribution along with a discriminative model that estimates whether the result is generated or not. However, the generator does not guarantee that each individual input image has a meaningful pair output image. For this reason, Zhu et al. [6] propose cycle consistent loss, in the sense that if we stylize, e.g., a real-world photo to a Monet-style painting, and then stylize it back, we should get the original photo. While fast and promising, these approaches suffer from the problem that the generative model is limited in 2D space and thus is not able to be conscious of the consistency in 3D space which is crucial for novel view synthesis. For this reason, we do not adopt the generative model-based method as our backbone.

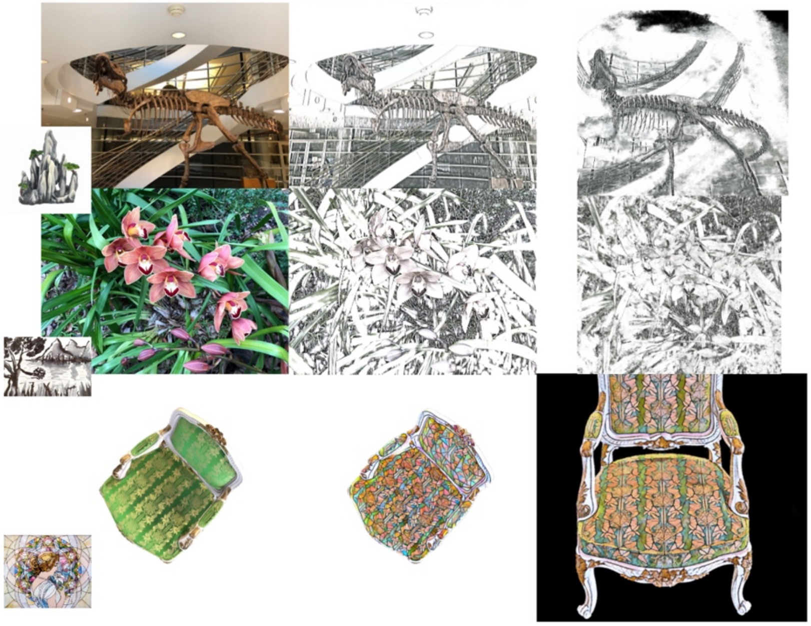

Figure 1: Our Method can generate all kinds of stylized images for both real-world datasets and synthetic datasets. The leftmost image in each row is the style image and original image. The center image is the stylized image. The rightmost image is the stylized novel view image.

Implicit 3D scene representations. Recently, implicit representations of complex 3D geometry shapes and textures have achieved significant popularity in learning-based 3D reconstruction. Mildenhall et al. [7] propose to use 2 inputs (spatial location and viewing direction) along with a multi-layer perceptron to represent a complex scene. It achieves state-of-the-art results in synthesizing novel views of a real-world scene with complex geometry and appearance. However, their method is restricted to real-world or synthetic scenes and objects. Inspired by the compelling result, we thus extend this work to non-photo-realistic images and demonstrate that NeRF can perform well on both real-world images and stylized images.

3. View-consistent Style Transfer

Given a set of real-world images taken from different viewpoints, our convolutional network translates them into a set of artistic images with an arbitrary style. This is achieved by deploying an efficient temporal loss item in the training phase.

In a pre-trained convolutional neural network e.g., VGG-16, filters in different layers can efficiently capture the corresponding feature of an input image [8]. A layer with n distinct filters has n feature maps of size m, where m is the height times the width of the feature maps. Those feature maps can be stored in a matrix $F^l \in R^{n*m}$, where F_{ij}^l is the activation value of the i^{th} filter at the flattened pixel position j in layer l. Then the content loss function can be defined as

$$L_{content}(p,x,l) = \frac{1}{2}\sum_{i,j}(P_{ij}^l - X_{ij}^l)^2 \tag{1}$$

where p and x are the original content image and the generated image. P_{ij}^l and X_{ij}^l are the corresponding feature matrix defined above. The hierarchy architecture of the convolutional neural network naturally possesses the ability that filters in the shallow layer tend to capture detailed information and those in the deep layer tend to capture high-level information.

Style information can be seen as the combination of activated filters in certain chosen layers. For this reason, it can be represented by the correlation of the feature matrix:

$$G_{ij}^l = \sum_k P_{ik}^l * P_{jk}^l \tag{2}$$

where G_{ij}^l is the inner product of the flattened feature maps i and j in layer l.

Let G_{ij}^l and H_{ij}^l be the respective gram matrix in layer l. The contribution of layer l can be defined as

$$S_l = \frac{1}{4n_l^2 m_l^2}\sum_{i,j}(G_{ij}^l - H_{ij}^l)^2 \tag{3}$$

then the total style loss is

$$L_{style}(p,x) = \sum_{l=0}^{L}\omega_l S_l \tag{4}$$

where L is the specific layer we choose for feature extraction.

To further address the view-inconsistency issue which appears in multi-view style transfer, we also introduce the view-consistency

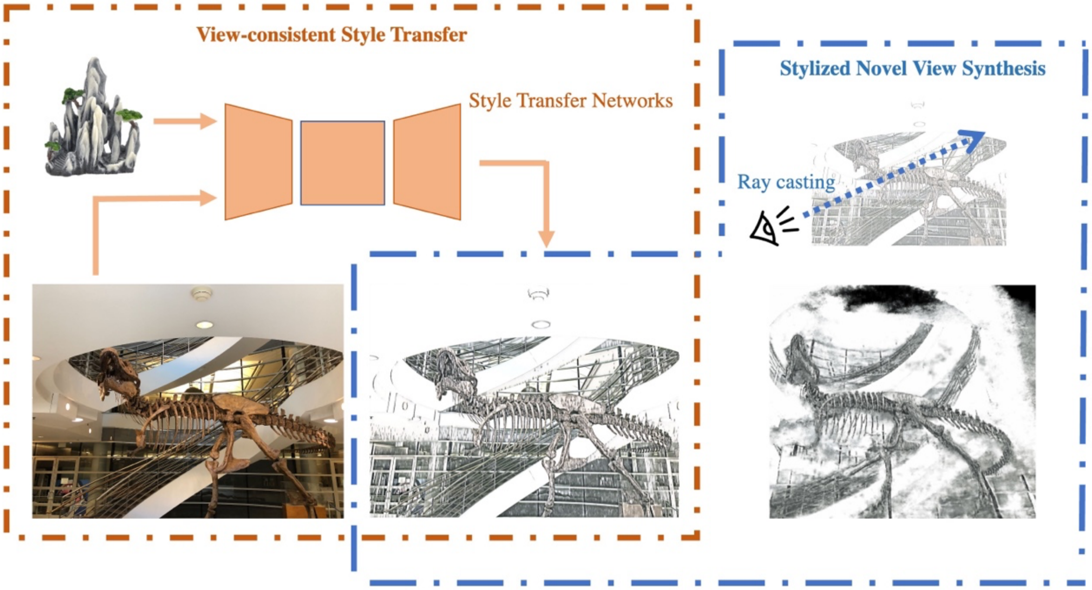

Figure 2: Overview of our proposed method. We first train a convolutional neural network (left) to stylize the original image to a certain style. We then use a radiance field (right) to encode the stylized complex scene. Once the training is done, we can synthesize the novel stylized views of the given scene.

loss, which is defined as

$$L_{consistency}(p,p_n)=\frac{1}{2}\sum_{i,j}(F(p)-F(p_n))^2 \tag{5}$$

where $F(\cdot)$ represents the feed-forward neural networks we are going to train. We add Gaussian noise to the original image p to get p_n. Then we compute the pixel-wised loss to constrain our model to the exact pixel values of the generated images despite the invariant noise.

We are inspired by our previous experiment where the original loss function performs poorly on video style transfer tasks. In that experiment, we shot a video (30fps) around a mountain model, and then we stylized the video frame by frame using the model trained without a view-consistency item. The stylized video turned out to be extremely inconsistent and considerable artifacts were generated. To better demonstrate the poor 3D-aware ability of the original model, we further performed dense 3D reconstruction using Structure from Motion (SfM). Reconstruction accuracy can be regarded as an important principle for multi-view consistency of the stylized images since high-fidelity reconstruction depends on consistency across views. As a result, we add the view-consistency loss item in the training phase to encourage the model to ignore the tiny change led by the movement of the camera and thus improved the model's robustness.

We train our neural network to minimize the loss function

$$L_{total}=\alpha L_{content}+\beta L_{style}+\gamma L_{consistency} \tag{6}$$

where α, β and γ are the weighting factor for content, style and view-consistency reconstruction.

4. Stylized novel view synthesis

In this section, we carefully describe how to synthesize stylized novel views in detail. Following Mildenhall et al. work, we implicitly represent a stylized scene using radiance field. The radiance field can be regarded as a continuous function that maps special location (x,y,z) and viewing direction (θ,φ) to volume density σ and RGB color c:

$$MLP(x,y,z,\theta,\varphi)\rightarrow(RGB\sigma) \tag{7}$$

where MLP is a multi-layer perceptron with learnable parameters. After that, we use traditional volume rendering techniques [9] to integrate these RGB values into an image. The expected RGB values are computed by the function:

$$C(r)=\int_{t_n}^{t_f}T(t)\sigma\big(r(t)\big)c\big(r(t),d\big)dt \tag{8}$$

$$T(t)=\exp\left(-\int_{t_n}^{t}\sigma\big(r(s)\big)ds\right) \tag{9}$$

Finally, we compute the MSE loss between the predicted color and ground truth color, and then update the parameters in MLP by backpropagation. The whole process is illustrated in Figure 2.

5. Experiment

Dataset: We conduct extensive experiments on both synthetic (Lego, Drum, Ship) and real-world (Fern, Flower, Horn, Orchid, Trex) datasets [7]. In addition, we further experiment with the mountains datasets that were created by ourselves. All scenes contain complex structures and delicate textures that are difficult to be reconstructed with conventional methods. We also test our method on diverse collections of artistic style images including Chinese ink wash painting, Claude Monet's Impression, Sunrise,

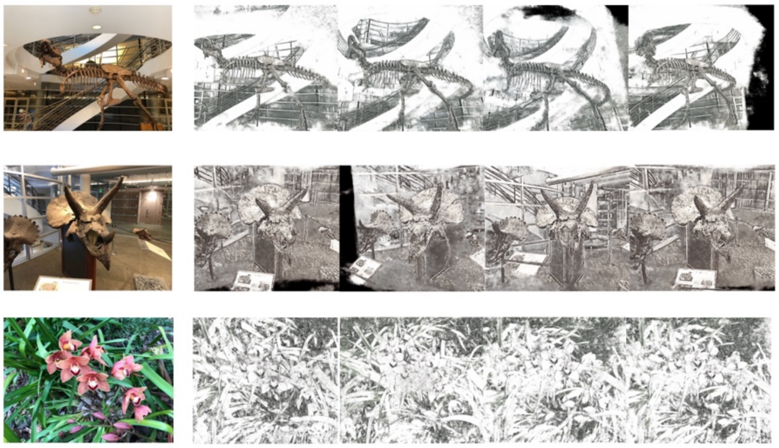

Figure 3：Our proposed method can synthesize view-consistent stylized free-view points. The leftmost image is the original image. The rest of them are stylized free-viewpoints images generated by the MLP.

Van Gogh's The Starry Night, mosaic, etc., to demonstrate our model's robustness over enormous ranges of artistic paintings.

Training Details: We follow Johnson et al. [3] work and train our style transfer networks on the Microsoft COCO dataset [10] with only two epochs. We resize the training images to 512×512 and set the batch size to 4. We apply Adam [11] optimizer with a learning rate of 1×10^{-3} and without using an exponential decay factor. We use relu2_2 to extract the content information and relu1_2, relu2_2, relu3_3, relu4_3 of the pre-trained VGG-16 to extract the style information. We find the model performs well on stylizing medium size images (image size from 256^2 to 1024^2). In addition, we set the α,β,γ equal to 1×10^6, 1×10^{11}, 1×10^2 respectively since we consider this to be a decent combination based on comparative experiments. We use PyTorch as our deep learning framework; training takes roughly 6 hours on a single GTX 3080Ti GPU. For the novel view synthesis, we apply instant-ngp [12] as our backbone since it can render the whole complex scene within 20 seconds.

Generated Free-Viewpoint Images: As shown in Figure 3, we perform ink wash style transfer tasks on 3 different datasets. The results show that the proposed method keeps the original object's style as well as content consistency with respect to the change of viewing direction.

6. Conclusion

We have introduced a method to synthesize artistic novel views of arbitrary scenes or objects using radiance fields. We extend the traditional style transfer task to dimensions beyond the 2D image plane, facilitating the creation of artistic work in 3D space. This research exhibits the promising capability of combining style transfer with implicit 3D representation.

References

[1] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[2] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[3] Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." *European conference on computer vision*. Springer, Cham, 2016.

[4] Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems* 27 (2014).

[5] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434* (2015).

[6] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017.

[7] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *European conference on computer vision*. Springer, Cham, 2020.

[8] Mahendran, Aravindh, and Andrea Vedaldi. "Understanding deep image representations by inverting them." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[9] Kajiya, James T., and Brian P. Von Herzen. "Ray tracing volume densities." *ACM SIGGRAPH computer graphics* 18.3 (1984): 165-174.

[10] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *European conference on computer vision*. Springer, Cham, 2014.

[11] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[12] Müller, Thomas, et al. "Instant neural graphics primitives with a multiresolution hash encoding." *arXiv preprint arXiv:2201.05989* (2022).