Transparent Ink-wash Style for Free-Viewpoint Video Generation

Zhizheng Xiang [1*], Hidehiko Shishido [1] and Itaru Kitahara [1]

[1] *University of Tsukuba, 1 Chome-1-1 Tennodai, Tsukuba, Ibaraki, Japan*

Abstract

We propose a framework that combines free-viewpoint video generation based on Neural Radiance Field (NeRF) with 2D ink-wash style transfer. In this work, we focus on 1) tackling the inconsistency issue caused by image style transfer and 2) synthesizing novel as well as stylized views of arbitrary given objects using NeRF, and 3) adding the transparent effect to ensure generated video looks more vivid.

Keywords

novel view synthesis, image style transfer, ink wash painting

1. Introduction

Deep learning-based artwork generation is a current trend in the world of art and is gaining popularity day by day. Compared to the traditional pipeline which involves a tremendous amount of time for a human expert, it requires little understanding of professional image editing skills and enables everyone to take part in the art generation community. Lots of traditional image editing company has also shifted to AI-generation technology. Adobe, for example, has been experimenting with generative AI for a long time and incorporated it into the software which empowers lots of creators.

However, most of the artworks generated by machine learning algorithms are limited to 2D space and perform unsuccessfully when applied to view-dependent image style transfer in 3D space due to appearance inconsistency issues. Extending style transfer techniques to dimensions beyond the 2D image plane introduces serious challenges and constraints. The core problem is how to preserve style consistency across the view to ensure high-quality view-consistent free-viewpoint video generation.

We propose an efficient style transfer pipeline for conducting style transfer tasks in 3D space. Our method can generate stylized ink wash free-viewpoint videos from multiple photos taken from different viewpoints. Our method can generate high-quality and vivid ink-wash-style scenes by adding a transparent effect, as shown in Figure 1.

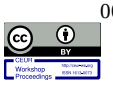
Our pipeline combines 2D image style transfer and 3D novel view synthesis and has high flexibility in choosing the style transfer and novel view synthesis models (Figure 2). Firstly, we run COLMAP [1] over a set of real-world images to calculate the camera position since the results are more accurate before stylization. Secondly, we treat a set of images taken from different viewpoints as a traditional style transfer task with the addition of a novel view consistency constraint, which aims to tackle the color inconsistency issue led by the classic image style transfer technique. Combining camera position with stylized images, we finally input that information into the Neural Radiance Field (NeRF) to train a multi-layer perceptron (MLP) and encode the stylized scene into the MLP for free-viewpoint video generation.

APMAR'22: The 14th Asia-PacificWorkshop on Mixed and Augmented Reality, Dec. 02-03, 2022, Yokohama, Japan

*Corresponding author.

EMAIL: xiang.zhizheng@image.iit.tsukuba.ac.jp (Zhizheng Xiang); shishido.hidehiko@image.iit.tsukuba.ac.jp (Hidehiko Shishido); kitahara@ccs.tsukuba.ac.jp (Itaru Kitahara)

ORCID: 0000-0003-0061-0530 (Zhizheng Xiang); 0000-0001-8575-0617 (Hidehiko Shishido); 0000-0002-5186-789X (Itaru Kitahara)



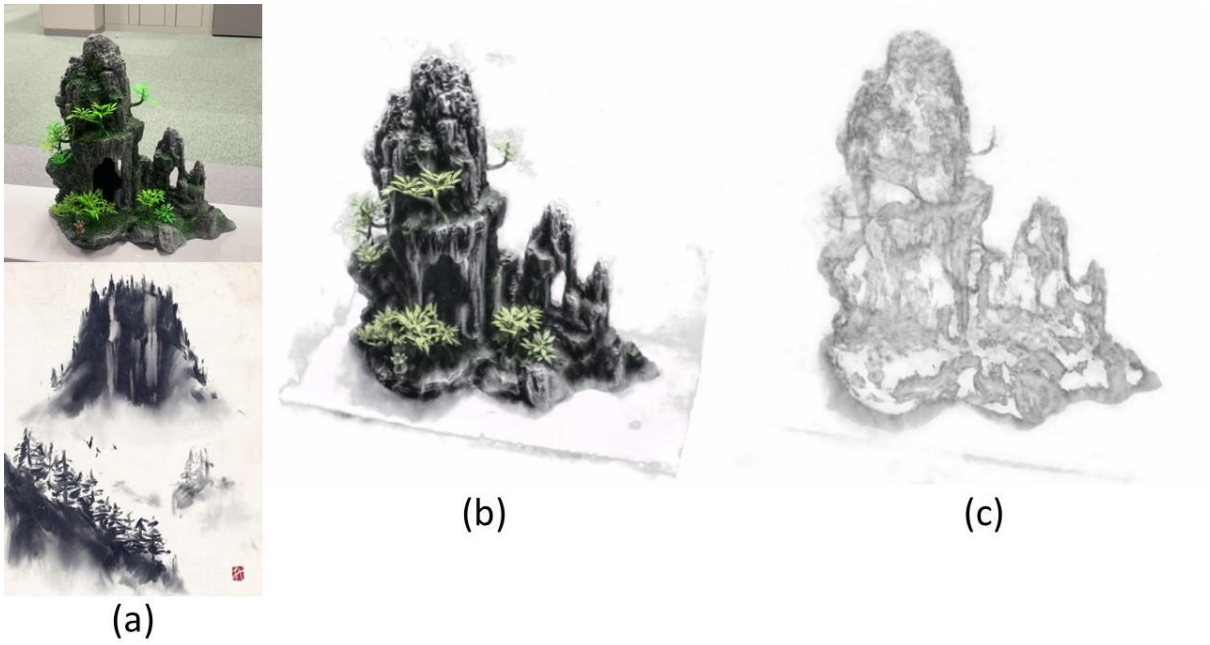


Figure 1: (a) shows the content (top) and style (down) image. (b) is a single frame of free-viewpoint ink-wash painting generated by our model without a transparent effect. (c) shows the result with a transparent effect.

We demonstrate that our style transfer model can be chosen from almost all types of ubiquitous models like feature extractor-based style transfer and generative model-based style transfer. On the other hand, our novel view synthesis model can be chosen from both traditional computer vision techniques and state-of-the-art volume rendering-based neural radiance fields.

In summary, our contributions are:

1. A highly efficient pipeline for stylizing arbitrary and complex 3D scenes that decently transfers the ink wash style details and maintains the view consistency with high visual quality.
2. A highly flexible and adaptive pipeline that is composed of 2D style transfer and 3D free-viewpoint generation.
3. We add a transparent effect to imitate the ink wash visual appearance which makes the generated results look more vivid.

2. Related works

Image style transfer. Currently, few types of models can fulfill image style transfer tasks at the human expert level. Gatys et al. [2] capture the style as well as content features by using a pre-trained convolutional neural network (VGG-16 or VGG-19 [3]) and then compute the loss between the content, style, and generated image to create entertaining artwork. Despite the high quality of the generated image, this method suffers from intolerable low speed (single image per minute). The later work proposed by Johnson et al. [4] overcomes the problem by training a feed-forward convolutional neural network to instantly approximate solutions to the neural style transfer problem.

Except for the feature extractor-based style transfer techniques we mentioned above, the generative model-based [5] style transfer techniques also appear to be attractive and promising in recent years. Radford et al. introduce Deep Convolutional GAN (DCGANs) [6] which can learn the hierarchy of representations from content image parts to style images. The main problem is that the generator does not guarantee that each input image has a meaningful pair of outputs since the whole task is an unsupervised learning task. Aiming to address this problem, Zhu et al. propose CycleGANs [7] implemented with cycle consistent loss. They first convert the image from the content set using a pair of the generative net and adversarial net and then revert it to the style set with another pair of neural networks. This method tends to train a more stable generator. In the field of ink-wash painting style transfer, He et al. [8] propose ChipGAN, a generative adversarial network for stylizing scenes in the Chinese ink-wash painting style. While fast and promising, the generative model-based style transfer is highly unstable when

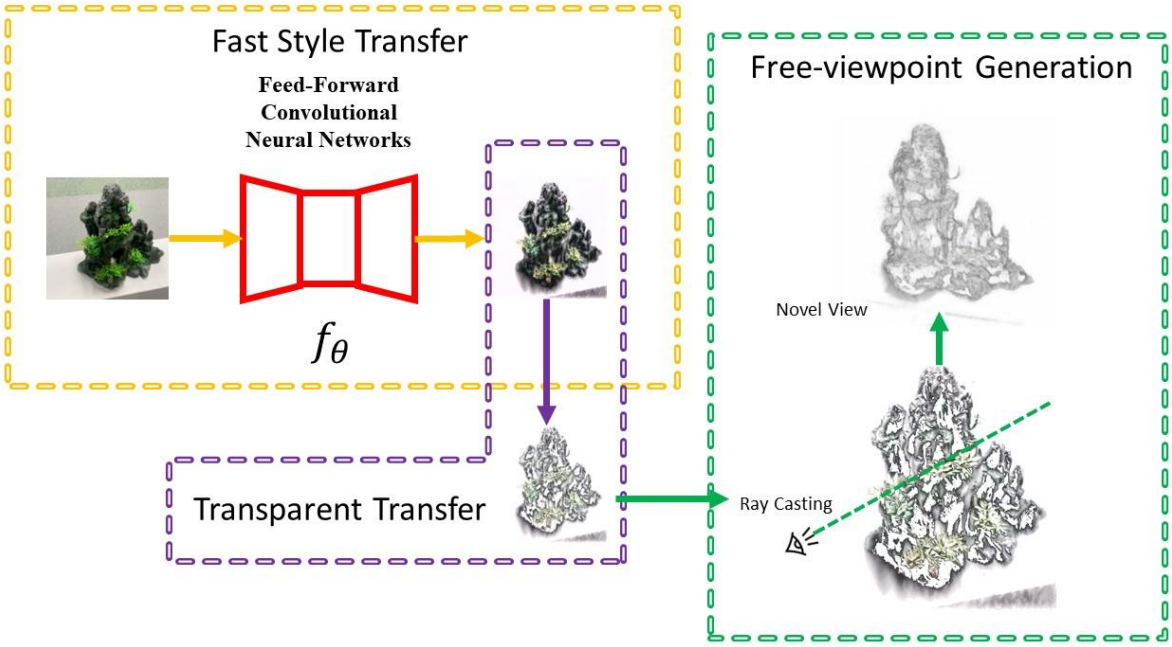


Figure 2: Overview of our pipeline. We first conduct style transfer to a set of real-world images (yellow part). Noted that we use a CNN in our experiment, but it can be replaced by other types of image style transfer techniques like GAN and VAE. Secondly, we add an additional alpha channel to enhance the visual quality of generated ink-wash paintings (purple part). Lastly, we encode our scene into a radiance field and use traditional volume rendering to compute pixel values to generate free-viewpoint video (green part). We use Instant-NGP for its high training and rendering speed but other types of NeRF models can also be used.

extend from a 2D image plane to 3D multi-view images. For this reason, we adopt feature extractor-based instead of generative model-based style transfer as our backbone.

Free-viewpoint generation. Structure from Motion (SfM), one of the most authoritative and traditional methods used for generating free-viewpoint video, has been studied for a long time and is still a very active research field in computer vision. Despite the accuracy of the reconstructed 3D structure, it struggles with synthesizing the interaction with light, especially the reflection effect. More importantly, SfM performs extremely awful in reconstructing a set of stylized images since it needs accurate cross-view feature points to match the geometric verification.

Recently, the implicit representation of complex scenes has retained a significant reputation in learning-based free-viewpoint video generation. Mildenhall et al. [9] propose to use spatial location and viewing direction to encode an arbitrary complex scene into a multi-layer perceptron and use the traditional volume rendering technique to compute the pixel values. It outperforms traditional 3D reconstruction methods in reconstructing free-viewpoint scenes with complex geometry and colorful appearance. Above all, implicit representation is highly suitable for reconstructing with a set of stylized images since it uses a neural network to describe scenes instead of explicit corresponding feature points. For this reason, we adopt implicit representation as our backbone for free-viewpoints generation.

3. Ink-wash painting style transfer

This section describes our view-consistent ink-wash painting style transfer method in detail. Given a set of real-world images taken from different viewpoints, our approach transfers them into a set of ink-wash paintings. We achieve this by implementing the fast style transfer method [4] (Sec. 3.1). We also introduce a novel view consistency loss that maintains the corresponding points' color as the viewpoint changes (Sec. 3.2). We add a transparent effect into the final results to further enhance our visual quality (Sec. 3.3).

3.1. Fast style transfer

Pretrained neural network has demonstrated their power in object detection, recognition, and

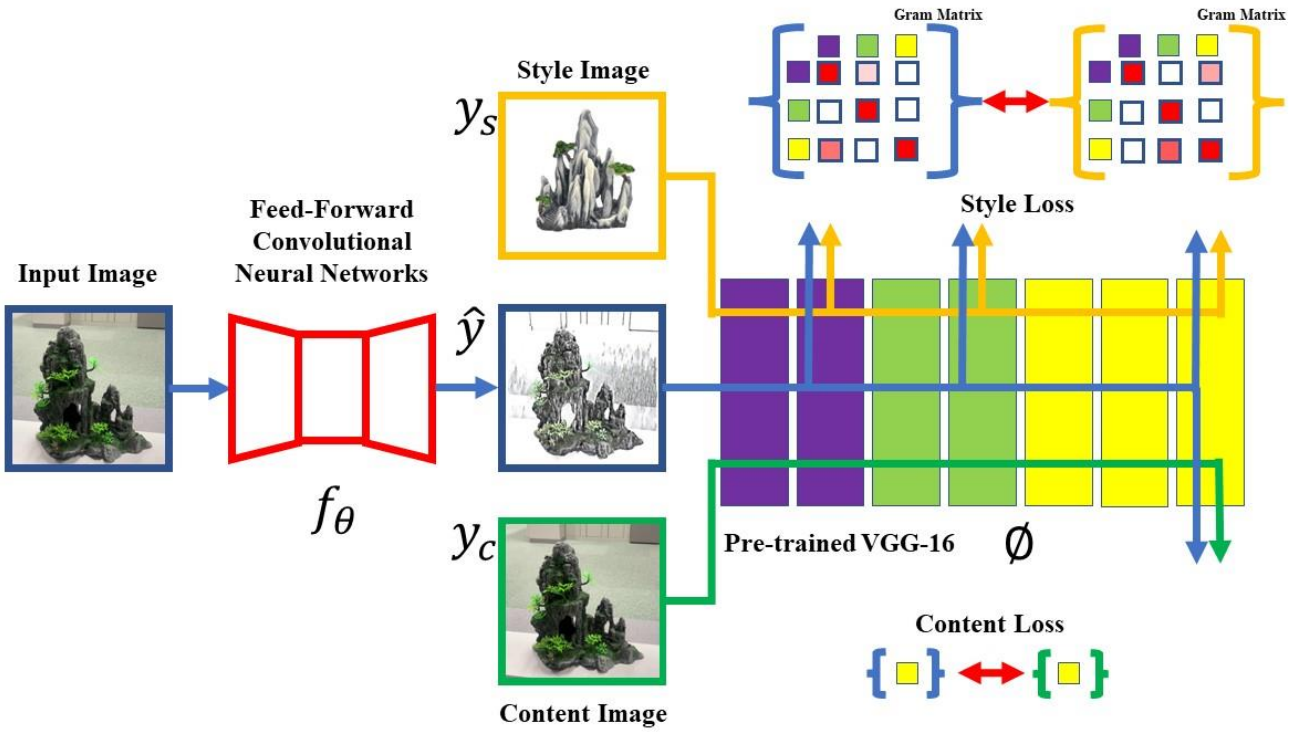


Figure 3: Overview of the fast style transfer. We train a feed forward CNN to stylize any image into ink-wash painting. We use pre-trained VGG-16 to capture the content and style features.

so on. Mahendran et al. [10] show that several layers in a pre-trained convolutional neural network retain the ability to represent high-level information and thus are eminently suitable for capturing both content and style features in an arbitrary image. As a result, Gatys et al. [2] minimize the content and style loss based on features extracted from a pre-trained VGG-16. Johnson et al. [4] train a feed-forward neural network and use it for style transfer tasks which is two orders of magnitude faster than Gatys's method.

In this experiment, we follow the architectural guidelines set by Johnson and Gatys. We train a feed-forward convolutional neural network to stylize input images into ink-wash paintings and use a pre-trained VGG-16 network to define content and style loss functions.

As shown in Figure 3, we first input the original content image x into our feed-forward convolutional network f_θ to generate the ink wash painting $\hat{y}$. Noted that one can also use a noise image with Gaussian distribution as the initial image, but it will generally take more training steps to converge. After stylizing, we input $\hat{y}$ paired with content image y_c into a pre-trained VGG-16 network to compute the content loss. A layer in VGG-16 has n distinct convolutional filters with 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 $\emptyset(*) \in R^{n*m}$, where $\emptyset$ is the pre-trained VGG-16 network and $\emptyset(*)_i$ is the activation value of the i^{th} flattened filter. Then the content loss function can be defined as

$$l_c(\hat{y}, y_c) = \frac{1}{2}\sum(\emptyset(\hat{y}) - \emptyset(y_c))^2 \qquad (1)$$

where $\hat{y}$ and y_c are the generated image and content image.

Style information, on the other hand, can be seen as the certain combination of activated filters in some chosen layers. It can be represented by the gram matrix:

$$G_{ij}(*) = \emptyset(*)_i * \emptyset(*)_j \qquad (2)$$

where G_{ij} is the inner product of the flattened feature maps $\emptyset(*)_i$ and $\emptyset(*)_j$.

Then we can define our style loss as:

$$l_s(\hat{y}, y_s) = \frac{1}{2}\sum(G_{ij}(\hat{y}) - G_{ij}(y_s))^2 \qquad (3)$$

where y_s is the style image.

Finally, we train our feed-forward neural network to minimize the total loss function:

$$l_{total} = \alpha l_c + \beta l_s \quad (4)$$

where α, β are the weighting factor for balancing the content and style strength.

3.2. View consistent constraint

We find that only performing vanilla style transfer to a set of real-world image sets will result in a noisy and unstable appearance even with little rotation or movement of the camera position. The main reason is that the feature extractor VGG-16 is trained on ImageNet [11] and all the images are restricted to a 2D image plane. 3D information such as time (in the case of a moving object), or 3D geometric shape is ignored. As a result, pre-trained VGG-16 can hardly possess the ability to realize the corresponding relationship of the same point if the viewpoint is changed.

This limitation motivates us to invent a novel view consistency loss that improves the visual consistency of stylizing images taken from different viewpoints. We first add Gaussian noise into input images and then compare the pixel-wised loss to constrain our neural network to the exact pixel values of the generated images despite the invariant noise. By calculating the view-consistency loss item along with content and style loss, we encourage the model to ignore the tiny change led by the tiny movement of the camera and thus improved the model's robustness.

We define our view-consistency loss as:

$$l_{consis}(x, x_n) = \frac{1}{2}\sum (f_\theta(x) - f_\theta(x_n))^2 \quad (5)$$

where x, x_n represents the input image and image with Gaussian noise separately. $f_\theta(*)$ represents the feed-forward neural network with parameters θ to be optimized.

In summary, the total loss function can be written as:

$$l_{total} = \alpha l_c + \beta l_s + \gamma l_{consis} \quad (6)$$

3.3. Transparent effect

Ink-wash painting is different from western art because it is an expression of spiritual contemplation. The content is more of abstract and smooth than observed effects and concrete natural details. Painters only use brush and ink to create the ink-wash painting thus the color shade becomes extremely important for an excellent painting. In our research, we aim to bring this effect into the AI art generation and make our final results look more vivid. To implement the transparent effect, we first convert our images file from RGB to CMYK using:

$$K = 1 - \max(R, G, B)/255 \quad (7)$$

RGB is the traditional file format to save the digital image while CMYK format is often used for printing. We can use 'K' values in the CMYK file to represent transparency degree since the generated images are mainly ink-wash style paintings. Finally, we concatenate 'K' value as an alpha channel to original RGB channel.

4. Free-viewpoint video generation

NeRF [9] proposes a neural radiance field to represent and reconstruct complex scenes implicitly, achieving state-of-the-art results in novel view synthesis. Following this work, we choose a branch of NeRF named Instant-NGP [12] for its fast training and rendering speed.

In general, the radiance field can be regarded as a continuous function that maps spatial location (x, y, z) and viewing direction (θ, φ) to volume density σ and RGB color c:

$$MLP(x, y, z, \theta, \varphi) \rightarrow (RGB\sigma) \quad (8)$$

where MLP is a multi-layer perceptron with learnable parameters.

After that, we apply traditional volume rendering techniques [13] to integrate these RGB values computed along rays into a pixel value. The expected RGB values are computed by the function:

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma\big(r(t)\big)c(r(t), d)dt \quad (9)$$

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma\big(r(s)\big)ds\right) \quad (10)$$

Finally, we compute the MSE loss between the predicted color and ground truth color, and then update the parameters in MLP by backpropagation.

Different from the original NeRF, our pipeline moves beyond photo-realistic rendering and combines ink-wash style with free-viewpoints video generation.

5. Experiments

In this section, we introduce the experiment datasets (Sec. 5.1) and then explain the details of our implementation of fast style transfer training process (Sec. 5.2). In the end, we show stylization results for various shapes of mountains (Sec. 5.3).

5.1. Mountain dataset

Mountain landscapes are by far the most prevalent topic in ink-wash painting. However, there are few datasets available that contain multi-view images of mountain models. As a result, we decide to buy some mountain models and create the dataset suiting for ink-wash painting style transfer task from scratch.

We create 3 different shapes of mountain models using iPhone 13 mini as our camera. The focal length is 26mm and all of the images' resolution is set to 1080×1080. Each dataset is composed of 70 to 100 images with spare different viewpoints and the camera pose is computed using COLMAP before stylization. Then we use the feed-forward neural network we trained before to stylize all of the mountain images. This process is extremely efficient and can be used for real-time stylization because we get rid of the computation of backpropagation. Lastly, we concatenate RGB value with an additional alpha channel which represents the transparency of a pixel.

5.2. Fast style transfer

We carefully follow Johnson et al. [4] work and train our fast style transfer networks on the Microsoft COCO dataset [14] with only two epochs because COCO dataset encompasses over 300, 000 images, and even only a single epoch is enough to learn the style mapping. We resize the training images to 512×512 and set the batch size to 4 because COCO dataset contains different kinds of objects and we consider a small batch size will guarantee the model learns the style mapping more efficiently. We apply Adam [15] optimizer with a learning rate of 1×10^{-3} and without using an exponential decay factor. We use relu2_2 to extract the content features and relu1_2, relu2_2, relu3_3, and relu4_3 of the pre-trained VGG-16 to extract the style features. 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 and training process 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.

Given a set of real-world images, our pipeline will generate a high-resolution free-viewpoints video within 1 minute if a pre-trained feed-forward style transfer neural network is used.

5.3. Results

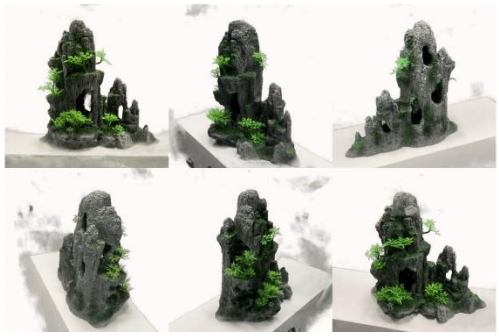


Figure 4: Free-viewpoints video generation of original mountain dataset

Figure 5: Free-viewpoints video generation of mountain dataset with ink-wash style. Style image is shown on the lower left side.

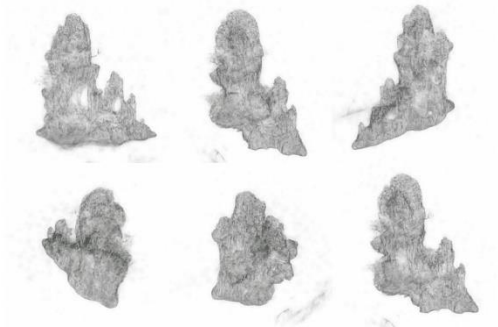


Figure 6: Free-viewpoints video generation of mountain dataset with ink-wash style and transparent effect.

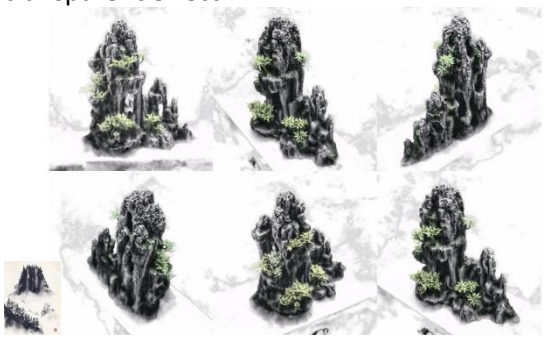


Figure 7: Free-viewpoints video generation of mountain dataset with another ink-wash style. Style image is shown on the lower left side.

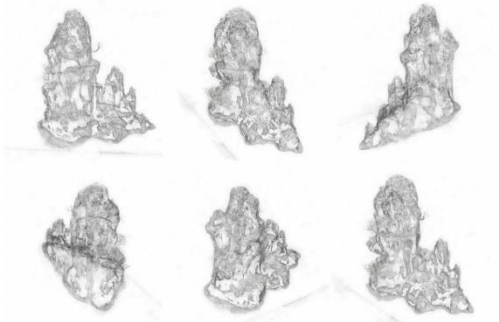


Figure 8: Free-viewpoints video generation of mountain dataset with another ink-wash style and transparent effect.

6. Conclusion

We have introduced a highly efficient pipeline to synthesize ink-wash painting style novel views of mountain scenes 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 2D image style transfer with implicit 3D representation.

7. References

[1] Schönberger, Johannes L., et al. "Pixelwise view selection for unstructured multi-view stereo." *European conference on computer vision*. Springer, Cham, 2016.

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

[4] 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.

[5] Goodfellow, Ian, et al. "Generative adversarial networks." *Communications of the ACM* 63.11 (2020): 139-144.

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

[7] 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.

[8] He, Bin, et al. "Chipgan: A generative adversarial network for Chinese ink wash painting style transfer." *Proceedings of the 26th ACM international conference on Multimedia*. 2018.

[9] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM* 65.1 (2021): 99-106.

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

[11] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." *International journal of computer vision* 115.3 (2015): 211-252.

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

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

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

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