GENERALIZABLE STYLE TRANSFER FOR IMPLICIT NEURAL REPRESENATION

Jaeho Moon **Taehong Moon** **Wonyong Seo**

ABSTRACT

Implicit Neural Representation (INR) has revolutionized scene representation using neural networks, offering high-quality rendering and memory efficiency. Leveraging a neural network as a continuous function, INR learns to render images through reconstruction loss, including RGB images for 3D scenes using multi-view images. In this project, we tackle the problem of style transfer for implicit neural representations. Existing methods for INR style transfer rely on test-time operations using fixed network parameters, leading to limited diversity in stylized images. To overcome these limitations, we propose an enhanced style transfer approach for INR. Our framework introduces generalizable INRs for style transfer, incorporating a ViT-based hypernetwork to predict instance-specific features for an MLP. By leveraging Adaptive Instance Normalization (AdaIN), we fuse content and style features, providing the MLP with the necessary inputs. We evaluate the effectiveness of our generalizable style transfer INR framework on the ImageNette dataset, demonstrating its ability to produce high-quality stylized images.

1 INTRODUCTION

Implicit Neural Representation (INR) has emerged as a novel approach for scene representation using neural networks. By treating a neural network, typically a Multi-Layer Perceptron (MLP), as a continuous function mapping pixel coordinates to color or features, INR learns to render images through reconstruction loss. Moreover, leveraging multi-view images and corresponding camera extrinsics, INR can render RGB images for 3D scenes (Mildenhall et al., 2021). With high-quality rendering and memory efficiency, INR offers a promising solution for scene representation.

Style transfer, a well-studied problem in computer vision, aims to apply a desired style to an image while preserving its underlying content structure. This project focuses on style transfer for implicit neural representations. Existing style transfer methods for INR typically involve test-time operations using the original content and style images, but they suffer from limited diversity in the resulting stylized images. Once the network parameters are optimized at test time, they remain fixed and generate the same stylized images for a given style.

To address these limitations, we propose an improved approach for style transfer in INR. We introduce a framework for style transfer of generalizable INRs. By adopting a ViT-based hypernetwork, we can predict instance-specific features to be given to an MLP. We utilize AdaIN proposed in (Huang & Belongie, 2017) to fuse content features and style features to be given to the MLP. With offline training of the framework, the MLP generates stylized images from unseen images without test-time optimization. Our generalizable style transfer INR framework can generate stylized images from unseen images on ImageNette dataset.

2 RELATED WORK

2.1 STYLE TRANSFER

The task of style transfer, which finds its roots in non-photo-realistic rendering and texture synthesis, has evolved significantly over time. Traditionally, style transfer involved formulating an optimization problem that simultaneously minimizes content and style loss based on the content and style images. While this approach has yielded satisfactory results, there is ongoing research to further improve the quality of style transfer.

One avenue of improvement lies in learning-based methods that leverage offline training on predefined datasets and utilize pre-trained networks. These methods offer greater flexibility and control over the style transfer process. For example, techniques such as adaptive instance normalization (AdaIN) have been introduced to align the mean and variance of content features with those of style features, facilitating more effective style transfer (Huang & Belongie, 2017). Other approaches explore reshuffling-based style loss (Gu et al., 2018) or employ whitening and coloring transforms (Li et al., 2017) to enhance stylization outcomes. Consistency in stylization without introducing unwanted artifacts has also been a focus of research (Li et al., 2018).

These learning-based methods build upon the foundation of traditional style transfer techniques, offering improved results and greater flexibility in controlling the style transfer process. By harnessing the power of pre-trained, frozen networks, these methods open up new avenues for creative expression and enable more sophisticated and refined style transfer applications. Ongoing research in this field continues to push the boundaries of what is possible, aiming to deliver even more compelling and realistic style transfer results.

2.2 Implicit Neural Representation (INRs)

Implicit Neural Representation (INR) is a powerful paradigm that leverages neural networks to capture intricate structures in continuous data, such as images, audio, and 3D objects. Initially, INRs utilized techniques like Fourier features and sinusoidal activations in coordinate-based MLPs to address spectral bias. These approaches have demonstrated impressive capabilities and have been applied in various domains, including data compression and rendering.

One notable application of INR is Neural Radiance Fields (NeRFs), which represent 3D scenes or objects as continuous functions. By training a neural network on multiple views of a scene, NeRFs can reconstruct the scene's geometry and render novel views from any camera position. While NeRFs excel at single-data representation, conventional INRs have limitations when it comes to accommodating multiple sets of data within a single network. Representing additional data points would typically require training a separate neural network from scratch.

To overcome this limitation, we propose a novel and generalizable INR framework. Our approach enables the representation of diverse datasets within a unified network, eliminating the need for training separate networks for each dataset. This advancement has significant implications for synthesizing and manipulating complex, multi-modal data. It provides a more efficient and scalable solution, opening up new possibilities for applications ranging from multi-domain style transfer to data-driven simulations. Our research contributes to the advancement of INRs, facilitating more versatile and flexible representations of continuous data.

2.3 Generalizable INRs

In order to address the challenge of representing diverse scenes without the need for training from scratch, researchers have introduced the concept of generalizable Implicit Neural Representations (INRs). These INRs are designed to learn modulated weights for neural networks, allowing them to be easily adapted to previously unseen data. Building upon the notion of modulation (Perez et al., 2018), alternative forms of INRs (Park et al., 2019; Mescheder et al., 2019) incorporate additional latent vectors that share weights across the dataset.

Taking inspiration from the concept of hypernetworks (Ha et al., 2016), several recent approaches (Dupont et al., 2022; Kim et al., 2022a; Szatkowski et al., 2022) leverage hypernetworks to predict modulated weights for Multi-layer Perceptrons (MLPs). This enables the extraction of common representations across the dataset. Notably, (Kim et al., 2022a) introduced a transformer-based hypernetwork specifically designed to predict modulated weights for coordinate-based MLPs, making them suitable for representing audio, images, and 3D views.

In our work, we adopt a transformer-based hypernetwork to predict weights for both content images and style images. By leveraging the capabilities of the hypernetwork, we aim to effectively capture the necessary information for style transfer while ensuring the flexibility and adaptability of the generalizable INRs.

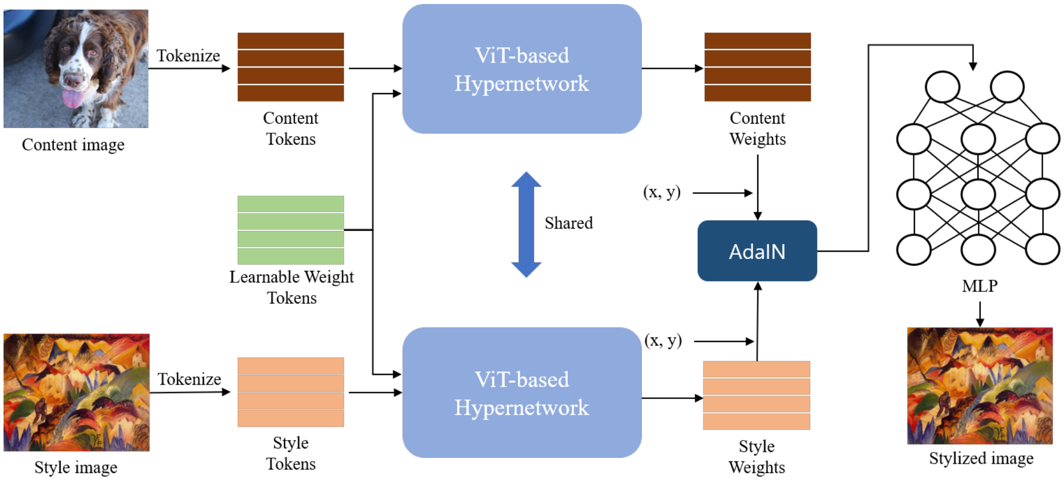

Figure 1: Our overall architecture.

3 Proposed Method

In this Section, we discuss the architecture design of our generalizable INR stye transfer and training scheme. In Section 3.1, we analyze the limitations of extending the previous approach (Kim et al., 2022b) for INR style transfer. Then, in Section 3.2, we describe the proposed architecture for generalizable INRs and how to train this network efficiently for stye transfer. In Section 3.3 we describe the training strategy of our proposed framework.

3.1 Motivating Experiments

The work by Kim et al. (2022) Kim et al. (2022b) introduced a style transfer approach for Implicit Neural Representations (INR) using a test-time training strategy. Their framework employs a Multi-Layer Perceptron (MLP) to generate stylized images based on a given content image and style image.

To achieve style transfer, they interpolate the latent vectors of the content image, denoted as z_c, and the style image, denoted as z_s, resulting in an interpolated latent vector $z' = \alpha z_c + (1-\alpha) z_s$, where α controls the degree of interpolation. This interpolated latent vector is then fed into the MLP along with positional encoded coordinates, and it encodes RGB values with single MLP for rendering. Compared with conventional INR training, the rendered image is penalized to ensure that the structure of the content image is preserved while transferring the style from the style image. During test-time adaptation, the level of stylization applied to the content image, which is decoded from INR, can be adjusted by α.

To generate the diverse stylized images, we first extend this framework in that it can encode both single content image and multiple style images at the same time. After training the provided architecture, we conducted an evaluation to assess its performance during test time when encountering unseen latent z inputs. In particular, the network was trained using a field image as the content image and a spring image as the style image. Subsequently, we evaluated the network by applying fall or winter styles while varying the alpha value. As depicted in Figure **??**, our findings indicate that the network exhibited deficiencies in accurately performing the desired stylization during test time and encountered reconstruction issues as well.

The observed outcomes can be viewed as a natural result, considering the reliance on pixel-based reconstruction loss during training. When presented with latent vectors that differ from the style images seen during training, it highlights a specific limitation of the encoded INR model in terms of its ability to generalize effectively.

In contrast, in this paper, we propose a style transfer framework based on generalizable INRs. Our approach aims to overcome the limitations of previous methods by developing a more versatile and adaptable style transfer system using INRs.

3.2 Proposed Architecture

The architecture of our proposed style transfer framework is depicted in Figure 1. It involves several steps to transform a given content image and style image into stylized outputs.

To begin, the content image and style image are tokenized into content tokens and style tokens using a Vision Transformer (ViT)-based hypernetwork (Dosovitskiy et al., 2020). This process allows us to extract meaningful representations from the input images. The ViT-based hypernetwork takes content tokens and learnable weight tokens as inputs and predicts content vectors. Similarly, it predicts style vectors for style tokens and learnable weight tokens.

Next, the content feature F_c and the style feature F_s are combined using the coordinates (x, y) and passed through an Adaptive Instance Normalization (AdaIN) layer (Huang & Belongie, 2017). The AdaIN layer applies a style transfer operation to match the statistics of the content feature with those of the style feature. It is expressed by the formula:

$$AdaIN(F_c, F_s) = \sigma_s \frac{F_c - \mu_c}{\sigma_c} + \mu_s, \tag{1}$$

where μ_c and μ_s represent the mean values of F_c and F_s respectively, and σ_c and σ_s represent their standard deviations.

The output of the AdaIN layer is then passed through a Multi-Layer Perceptron (MLP), which performs additional transformations and computations to generate the final stylized images. The MLP takes the combined features as input and produces the rendered stylized images as output.

By following this architecture, we are able to achieve style transfer by leveraging the content and style representations obtained from the ViT-based hypernetwork, and applying the AdaIN layer and MLP for the final stylized image generation process.

3.3 Training Strategy

To ensure the generalizability of Implicit Neural Representations (INRs), we employ a two-stage training process. In the first stage, we pretrain both the Vision Transformer (ViT)-based hypernetwork and the Multi-Layer Perceptron (MLP) simultaneously using the training dataset.

During this pretraining stage, we feed a single image as input to the network and pass it through the AdaIN layer. It is important to note that in this stage, the AdaIN layer operates with the same feature, meaning that the statistics of the feature are not altered by the AdaIN transformation. This helps to maintain consistency and stability during the pretraining process.

Once the pretraining stage is completed, we freeze the parameters of the ViT-based hypernetwork, as it has already learned to extract meaningful content and style representations. We then proceed to finetune the remaining components of the framework using both content images and style images.

During the finetuning stage, we train the network to optimize the style transfer performance by adjusting the parameters of the MLP. By using content images and style images together, the network learns to combine the content features and style features through the AdaIN layer, generating stylized outputs that preserve the structure of the content image while incorporating the style characteristics of the style image.

This two-stage training approach allows us to effectively train generalizable INRs by initially pretraining the network to learn meaningful representations and then fine-tuning it to specialize in style transfer tasks using content and style images.

3.4 Loss

Image reconstruction loss. At pertaining, we use image reconstruction loss to ensure model reconstructs given image well. For given image x_1, encoder E, and composed MLP INR, image reconstruction loss can be formulated as

$$\mathcal{L}_{recon} = \sum_{(i,j) \in H \times W} ||INR(i, j; E^c(x_1)) - x_1(i, j)|| \tag{2}$$

where $E(\cdot)$ is hypernetwork, $INR(\cdot; c)$ is composed MLP, and $x(i, j)$ are color value of image x at coordinate (i, j).

Content loss. At second stage, We also have to ensure context have to be maintained in stylized image $x_{c \mapsto s}$. We utilize VGG network to extract context information from content image and stylized image.

$$\mathcal{L}_{content} = ||VGG(x_{c \mapsto s}) - VGG(x_c)|| \tag{3}$$

Style loss. We additionally employ style loss to make stylized image have similar style with stylized image. We define style loss similar to Kim et al. (2022b), by measuring L2 norm between gram matrices $\mathcal{G}$ from style image and stylized image.

$$\mathcal{L}_{style} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}[||\mathcal{G}(x_{c \mapsto s}) - \mathcal{G}(x_s)||] \tag{4}$$

Total loss. Total loss is as follows:

$$\mathcal{L}_{stage1} = \mathcal{L}_{recon} \tag{5}$$

$$\mathcal{L}_{stage2} = \lambda_{content}\mathcal{L}_{content} + \lambda_{style}\mathcal{L}_{style} \tag{6}$$

where $\lambda_{content}$, λ_{style} are weights for each loss.

4 Experimental Results

4.1 Datasets

We utilize Imagenette(Howard) and WikiArt(Saleh & Elgammal (2015)) for universal style transfer. We first tried to train our model on full dataset, but the model took an excessively long time to converge. So, we randomly selected 1000 style images from WikiArt dataset and 1000 content images from Imagenette dataset for training.

4.2 Training details

We implement our model in Pytorch. We train our model with AdamKingma & Ba (2014) optimizer with cosine annealing schedulingLoshchilov & Hutter (2016) with max learning rate $1e^{-4}$ for 2000 epochs at first stage. At second stage, we additionally train the model with $\mathcal{L}_{stage2}$ for 500 epochs. We also use Adam optimizer, cosine annealing scheduling with max learning rate $1e^{-4}$. Also, we randomly select content and style image to ensure universal style transfer.

4.3 Qualitative Results

We show stylized image from our model and INR-st(Kim et al. (2022b)) in Figure.2. We can found that stylized images from our model contain fine structure of content images and color similarity from style images. INR-st also generates well-stylized images, but it generates unrealistic clouds (3rd row in Figure 2) and flowers (4th row in Figures 2). However, our results show blurrier boundaries. It might come from universal style transfer training of our model, while INR-st only optimized on one pair of content and style images. We also expect that our model could generate sharper stylized images when we found more precise hyperparameters and longer training steps.

5 Conclusion

In conclusion, we have presented an improved approach for style transfer in Implicit Neural Representations (INR), aiming to enhance the diversity and flexibility of stylized images. Our framework utilizes a ViT-based hypernetwork to predict instance-specific features for an MLP, enabling fine-grained control over the style transfer process. By incorporating Adaptive Instance Normalization

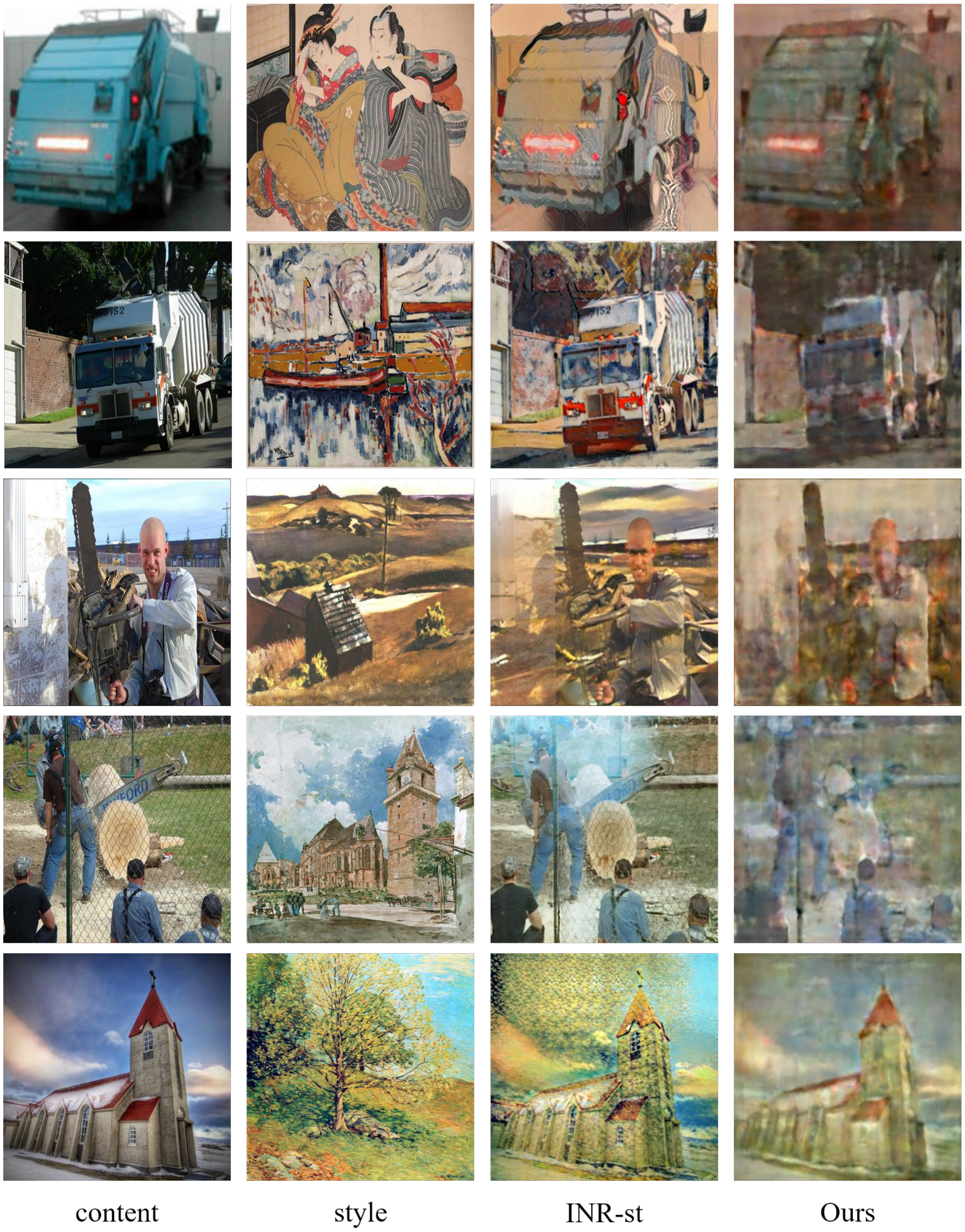

Figure 2: Qualitave comparison between our model and INR-st(Kim et al. (2022b)). Note that INR-st is optimized on one pair of content and style image, while our model generates stylized images from various input content and style image pairs.

(AdaIN), we successfully fuse content and style features, providing the necessary inputs for generating stylized images. Experimental evaluations on the ImageNette dataset have demonstrated the feasibility of our generalizable style transfer INR framework in producing stylized images. In the future, we can extend this work by training arbitrary scale rendering to produce arbitrary scale style transfer by utilizing the characteristics of INRs.

REFERENCES

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Emilien Dupont, Hyunjik Kim, SM Eslami, Danilo Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you should treat it like one. *arXiv preprint arXiv:2201.12204*, 2022.

Shuyang Gu, Congliang Chen, Jing Liao, and Lu Yuan. Arbitrary style transfer with deep feature reshuffle. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8222–8231, 2018.

David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

Jeremy Howard. Imagewang. URL https://github.com/fastai/imagenette/.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Chiheon Kim, Doyup Lee, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Generalizable implicit neural representations via instance pattern composers. *arXiv preprint arXiv:2211.13223*, 2022a.

Sunwoo Kim, Youngjo Min, Younghun Jung, and Seungryong Kim. Controllable style transfer via test-time training of implicit neural representation. *arXiv preprint arXiv:2210.07762*, 2022b.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 453–468, 2018.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

Filip Szatkowski, Karol J Piczak, Przemysław Spurek, Jacek Tabor, and Tomasz Trzciński. Hypersound: Generating implicit neural representations of audio signals with hypernetworks. *arXiv preprint arXiv:2211.01839*, 2022.