Diffusion-Enhanced PatchMatch: A Framework for Arbitrary Style Transfer with Diffusion Models

Mark Hamazaspyan[1,2], Shant Navasardyan[1,3]
[1] Yerevan State University (YSU), [2] Metric, [3] Picsart AI Research (PAIR)

Abstract

Diffusion models have gained immense popularity in recent years due to their impressive ability to generate high-quality images. The opportunities that diffusion models provide are numerous, from text-to-image synthesis to image restoration and enhancement, as well as image compression and inpainting. However, expressing image style in words can be a challenging task, making it difficult for diffusion models to generate images with specific style without additional optimization techniques. In this paper, we present a novel method, Diffusion-Enhanced PatchMatch (DEPM), that leverages Stable Diffusion for style transfer without any finetuning or pretraining. DEPM captures high-level style features while preserving the fine-grained texture details of the original image. By enabling the transfer of arbitrary styles during inference, our approach makes the process more flexible and efficient. Moreover, its optimization-free nature makes it accessible to a wide range of users.

1. Introduction

Style transfer has emerged as a popular research area in computer vision [1, 6, 10, 12, 12, 19], enabling the synthesis of visually appealing images by combining the content of one image with the style of another. Broadly, style transfer techniques can be categorized into three types: optimization-based [6, 10], neural network training-based [12, 26], and arbitrary image style transfer [1, 19].

DEPM falls into the third category, which aims to provide a more flexible and efficient solution for transferring arbitrary styles without the need for training. Existing methods for arbitrary style transfer have their own advantages and disadvantages. For instance, optimization-based methods often produce high-quality results but are computationally expensive, while neural network training-based methods require extensive training for each specific style. Our method addresses these limitations by synergistically combining diffusion models with style transfer techniques, enabling the transfer of arbitrary styles during the inference of Stable Diffusion [22] without any finetuning or pretraining.

Recent works, such as those by Ho et al. [7] and Nichol et al. [20], have demonstrated the potential of diffusion models for a wide range of applications. Furthermore, researchers have explored the use of diffusion models for tasks beyond image synthesis, such as video synthesis [3, 15], audio synthesis [16], and even reinforcement learning [11]. These advances highlight the versatility and potential of diffusion models for various domains and encourage further exploration of their capabilities.

Figure 1. Style trainfer with DEPM, for each pair of images the content image is the left one with the style image in the top left corner of it, and the stylized image is on the right.

In this paper we propose a new approach for style transfer by leveraging the generative power of diffusion models. Some results can be found in Fig. 1.

Our contributions can be summarized as follows:

1. We utilize patch-based techniques with whitening and coloring transformations in the latent space of Stable Diffusion for high-quality arbitrary style transfer.

2. Our approach demonstrates superior performance in terms of color transformation while preserving the content details of the input image.

3. Our method enables arbitrary style transfer without the need for any training, making it a highly flexible and efficient solution for a wide range of applications.

2. Related work

2.1. Style transfer

The field of style transfer has seen significant advancements in computer vision research in recent years, with various approaches proposed to tackle this problem [2,9,17,28]. In this section, we review some of the most notable works in the area, highlighting their strengths and limitations.

One of the earliest and most widely cited works in this field is Gatys et al. [6], which introduced a neural network-based approach for style transfer. The approach leverages a pre-trained convolutional neural network to separate the content and style of an input image, before optimizing a new image that preserves the content of the input image while adopting the style of a target image. This method produces visually appealing results and has been used in many artistic applications. However, it is computationally expensive, requiring iterative optimization over a large image space, and it struggles to handle color transformation. To address these limitations, several works have proposed alternative optimization techniques, such as gradient descent or more efficient neural network architectures, such as feed-forward networks (Johnson et al., [12]). These approaches achieve faster inference times and have less computational cost, making them more practical for real-time applications. However, they often sacrifice the quality of the results for the sake of efficiency.

Another notable approach is introduced by Chen et al. [1], which uses patch matching to align the textures between the input and target style images. The method then transfers the aligned texture patches to the input image to create the final stylized image. This approach achieves impressive results in preserving local texture details and has a faster inference time than optimization-based methods. However, it struggles to handle global style transfer and can result in artifacts in the stylized image.

Whitening and Coloring Transform (WCT) by Li et al. [19] is another popular approach that uses a whitening operation to separate the content and style of an input image before transferring the style using a coloring operation. This method achieves high-quality results while preserving the content of the input image. However, it is not as flexible as other methods, as it requires pre-defined sets of styles for training.

In recent years, Exact Feature Distribution Matching for Arbitrary Style Transfer and Domain Generalization by Zhang et al. [30], proposed a novel approach for arbitrary style transfer that matches the distributions of features between the content and style images. This method achieves state-of-the-art results for arbitrary style transfer, while also being effective for domain generalization, enabling the model to generalize to previously unseen styles. However, it requires significant computational resources for training and inference.

In the light of feature distributions: moment matching for Neural Style Transfer by Kalischek et al. [13] also addresses the feature distribution matching problem in neural style transfer. By matching the first and second moments of feature maps from the content and style images, the method achieves high-quality stylization results. However, it is not as effective for arbitrary style transfer and may produce overly stylized images.

Domain-Aware Universal Style Transfer by Hong et al. [8] proposes a method for universal style transfer that can learn to transfer style across different domains. This method achieves superior results for universal style transfer, while also being flexible and efficient. However, it requires a large number of styles for training, which can be time-consuming and computationally expensive.

In summary, the style transfer field has seen significant progress in recent years, with various approaches proposed to tackle the problem. While each approach has its strengths and limitations, recent approaches, such as Exact Feature Distribution Matching for Arbitrary Style Transfer and Domain Generalization, Domain-Aware Universal Style Transfer, and feature distribution moment matching methods, offer promising avenues for more efficient, flexible, and effective style transfer.

2.2. Diffusion models

Diffusion models have emerged as a powerful class of generative models that learn to generate images by simulating a diffusion process [7,25,27]. These models consist of a series of noise-corrupted images, where the noise is gradually removed in a step-by-step manner. By learning this reverse diffusion process, diffusion models can generate high-quality images that capture the underlying data distribution. In this section, we provide an overview of diffusion models, their mathematical formulation, and their applications, including Textual Inversion, Dreambooth, and others.

2.2.1 Applications

Diffusion models have been successfully applied to a wide range of tasks, including image synthesis, inpainting, denoising, and more. Some notable applications and research areas are discussed below:

Textual Inversion: Textual inversion [5] is a technique that aims to generate images from textual descriptions by inverting the diffusion process. By conditioning the diffusion model on text embeddings, researchers have been able to generate images that closely match the given textual descriptions. This demonstrates the potential of diffusion models for text-to-image synthesis and their ability to capture complex semantic information.

Dreambooth: Dreambooth [24] is an interactive image synthesis application that allows users to generate images by providing textual descriptions and adjusting various parameters. By leveraging diffusion models, Dreambooth enables users to generate high-quality images with fine-grained control over the content and style. This showcases the flexibility and expressiveness of diffusion models for creative applications.

Style Transfer: Diffusion models can be combined with style transfer techniques to enable the transfer of arbitrary styles during the inference step without any finetuning or pretraining. This integration not only enhances the performance of style transfer but also allows for text-guided style transfer, providing users with greater control and customization over the generated output [18].

3. Method

Our approach synergistically combines diffusion models with style transfer techniques, leveraging the power of generative diffusion models to capture high-level style features while ensuring that the fine-grained texture details of the original image are preserved.

3.1. Background

3.1.1 Diffusion models

The diffusion process can be described as a Markov chain, where each step involves adding noise to an image. Given an initial image x_0, the diffusion process can be defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{2}$$

where $\sqrt{1-\beta_t}x_{t-1}$ and β_t are the mean and variance of the Gaussian distribution, respectively. The goal of diffusion models is to learn the parameters of this conditional distribution, which can be achieved using a denoising score matching objective. The reverse diffusion or denoising process is defined as:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{3}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{4}$$

One notable class of diffusion models is the Latent Diffusion models by Rombach [23]. Latent diffusion models simplify the diffusion process by projecting high-dimensional input into a smaller latent space. To achieve this, an encoder network is used to encode the input into

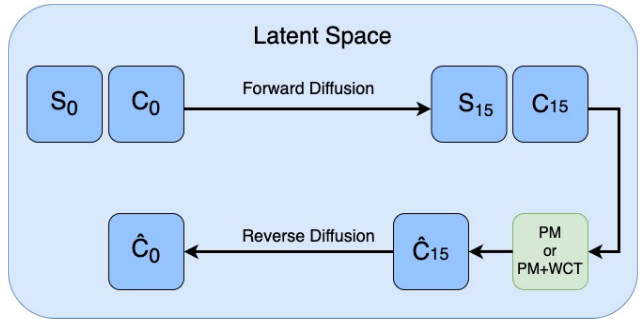


Figure 2. C_0 and S_0 are the latent representations of content and style images at timestamp $t = 0$. Initially, we perform $t = 15$ steps of deterministic forward pass with LMSDiscreteScheduler to get C_{15} and S_{15} (total number of steps is $T = 100$), then we start the reverse diffusion process. At $t = 15$ we perform PM or PM followed by WCT to get the latent representation $\hat{C}_{15}$. We finish the process by performing $t = 15$ reverse diffusion steps on $\hat{C}_{15}$ resulting in $\hat{C}_0$, the latent representation of the stylized image.

a latent representation, which is then processed by a standard diffusion model to generate new data. This reduces the computational demands of training diffusion models by processing the input in a lower dimensional space. The resulting data is then upsampled by a decoder network.

Stability AI has open-sourced the weights of its own Latent Diffusion model: Stable Diffusion which is used for the experiments in this paper.

3.1.2 Patch Match (PM)

The primary component of PM involves a patch-based operation that constructs target activations in a single layer, utilizing both style and content images [1]. The style swap procedure comprises the following steps:

1. Extract a collection of patches from both content and style activations.
2. For every content activation patch, identify the most closely-matching style patch using the normalized cross-correlation measure.
3. Replace each content activation patch with its nearest-matching style patch.
4. Reassemble the entire content activations by averaging overlapping regions that may possess different values due to step 3.

3.1.3 WCT

In the whitening step, the content image's feature maps are transformed to remove the original style information by making the covariance matrix of the content features close to an identity matrix. This is achieved by applying a decorrelation operation, which involves eigendecomposition of the content covariance matrix followed by rescaling the content features using the inverse square root of the eigenvalues. The coloring step, on the other hand, involves transferring the style information from the reference image to the whitened content features. This is accomplished by computing the covariance matrix of the style features and then applying a correlation operation, which involves multiplying the whitened content features with the square root of the style eigenvalues and the style eigenvectors. The resulting feature maps, which now possess the style of the reference image and the content structure of the original image, are then passed through a decoder network to generate the final stylized output [19].

3.2. Diffusion-Enhanced PatchMatch (DEPM)

In our experiments, we utilize the publicly available Stable Diffusion model[1]. We begin by performing $t = 15$ (the total number of steps is $T = 100$) steps of deterministic forward pass with LMSDiscreteScheduler [14] to C_0, the latent representation of the content image. The resulting latent code is hence denoted by C_{15}. For a 100-step inference process, we determine that 15 steps were optimal for content preservation. Similarly for the style image S_0, we get the latent representation S_{15}. Then, we apply the patch match transformation to the latent variables of the corresponding content and style images at step $t = 15$ resulting in a latent representation $\hat{C}_{15}$. Finally, we perform $t = 15$ backward diffusion steps on $\hat{C}_{15}$ (with the scheduler LMSDiscreteScheduler) resulting in the $\hat{C}_0$, latent representation of the stylized image. To get the stylized image it remains to pass $\hat{C}_0$ to the decoder of Stable Diffusion. The overview of our method is shown in Fig. 2.

In our experiments, we focused on the style transfer application and left the text prompt empty to concentrate solely on the transfer of artistic styles between images. This allowed us to thoroughly evaluate the performance of our method in terms of color transformation, artistic style transfer, and content preservation. Nevertheless, there is ample room for further research, where the textual prompt can be utilized for content image modification, combined with style transfer.

We explored various combinations of transformations, including PM only, WCT only, and PM followed by WCT. After implementing the PM, the transformed latent code's pixels do not exhibit a Gaussian distribution, leading to a denoising process that produces output from a distinct distribution compared to the input. Consequently, we employ WCT to bring the covariance matrix of the transformed latent code closer to the Identity matrix. This approach substantially enhances the results, as demonstrated in Fig. 4 of the Ablation study section. Moreover, we empirically came to the conclusion that by multiplying the style latent with constant $\sigma = 1.5886$ hyperparameter right before the PM transformation, the results are getting better compared with just PM. Our experiments indicate that multiplying the latent with sigma expands its distribution tails, which in turn facilitates the style transfer process by enabling a more effective transfer of the style characteristics from the style image to the content image. This observation highlights the importance of carefully tuning the hyperparameters in our method to achieve optimal style transfer results. For our experiments, we employed the LMSDiscreteScheduler [14] scheduler, but any other scheduler can be used for our technique. This approach demonstrates the flexibility and effectiveness of our method in achieving high-quality style transfer results while preserving the content details of the input image.

[1]Stable Diffusion weights https://huggingface.co/CompVis/stable-diffusion-v-1-4-original/tree/main.

4. Experiments

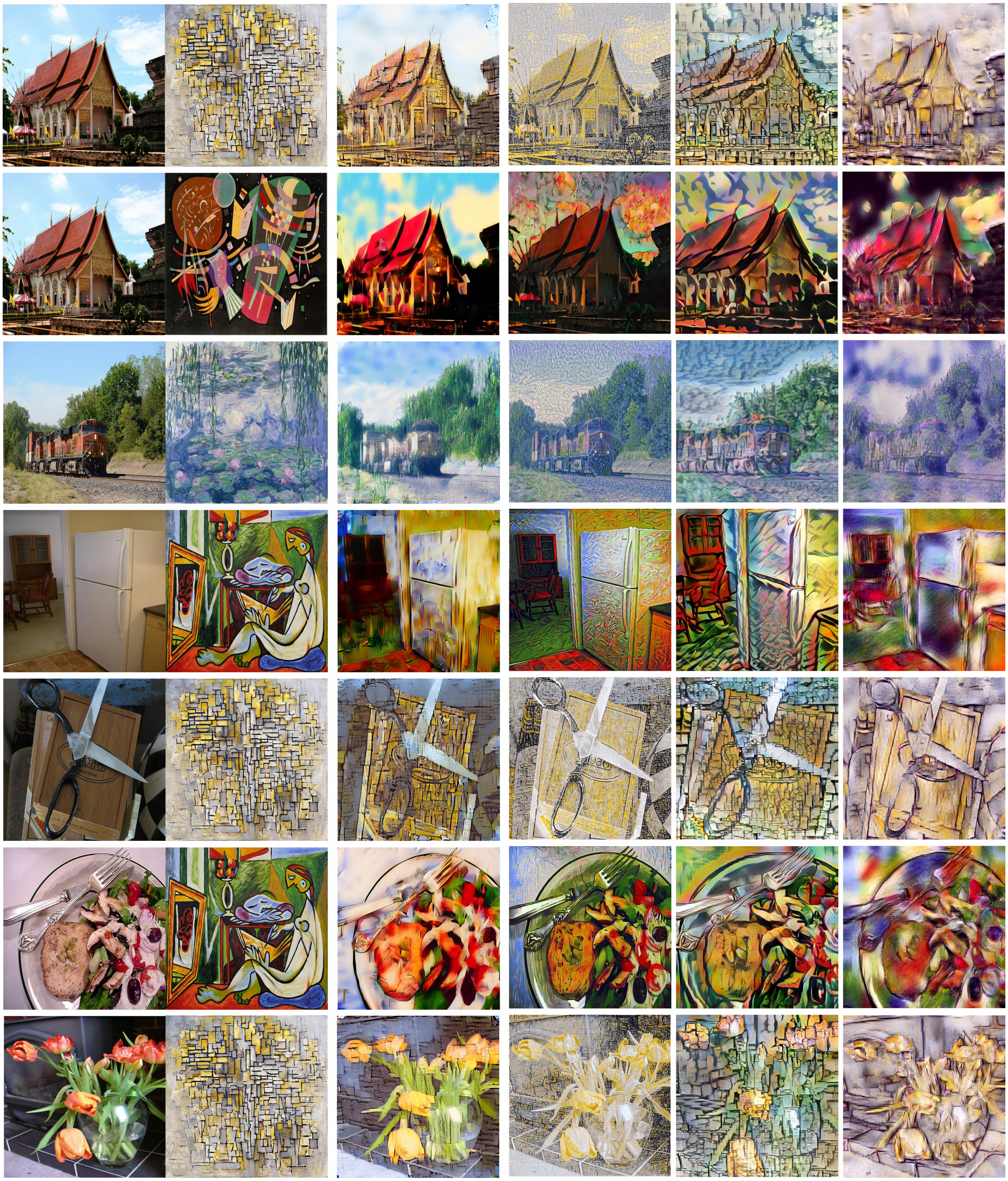
In this section, we present a series of experiments designed to evaluate the performance of our proposed method and compare it with three other state-of-the-art style transfer techniques. Our goal is to demonstrate the effectiveness of our approach in achieving high-quality style transfer results while preserving the content details of the input image. We provide a brief introduction to the experimental setup, followed by a detailed analysis of the results and comparisons with the selected methods. In this section, we provide a qualitative comparison of our proposed method with the three selected baseline methods: EFDM (Example-based Feature-driven Diffusion Model), CMD (In the light of feature distributions), and DSTN (Domain-Aware Universal Style Transfer). We focus on the ability of each method to transform colors and artistic styles, preserve content components, and distribute style effects across unique objects in the generated images.

4.1. Implementation details

The experiments were conducted on a diverse set of content and style images, covering various artistic styles and image content. We used images from PascalVOC [4] and WikiArt [21] datasets for style transfer. Throughout the experiments, we set 512x512 as the default image resolution. We set the LMSDiscreteScheduler to $T = 100$ timestamps and used $t = 15$ steps to add noise to the content image. During the denoising steps we perform PM and WCT transformations at step $t = 15$. Our code is implemented with PyTorch and inference is done on a single GeForce GTX 1080 Ti.

4.2. Qualitative comparison

For our experiments, we selected three representative style transfer methods from the literature to serve as baselines for comparison. These methods encompass a diverse



(a) Content (b) Style (c) **DEPM (Ours)** (d) CMD [13] (e) EFDM [30] (f) DSTN [8]

Figure 3. Comparison of different methods for image style transfer.

range of approaches. By comparing our method with these baselines, we aim to showcase the advantages of our approach in terms of flexibility and efficiency. Fig. 3 shows the various content and style images (columns 1,2 respectively), our approach (column 3), CMD (column 4), EFDM (column 5) and DSTN (column 6). One can see that our approach works stably on transfering colors and artistic elements of style images.

Our approach demonstrates a superior performance in terms of color transformation and artistic style transfer. Compared to EFDM, our method better preserves the content components of the content image, ensuring that the fine-grained details and structural information are maintained throughout the style transfer process. This results in synthesized images that exhibit a more accurate and visually appealing transfer of style while maintaining the integrity of the original content. Rows 4,5 and 7 are expressive examples of the above points.

When compared to CMD, our method produces images with more well-defined artistic components. The overall style is not uniformly distributed across the image, allowing for a more nuanced and context-aware transfer of style that better affects unique objects within the content image (more expressive in Row 1,4,7). This leads to a more engaging and visually striking output that effectively captures the essence of both the content and style images.

Finally, in comparison to DSTN, our approach performs better in both artistic style transfer and content preservation. By synergistically combining diffusion models with style transfer techniques, we are able to achieve a more accurate and visually appealing transfer of style while ensuring that the content details of the input image are preserved.

In summary, our method demonstrates a strong qualitative performance in comparison to the selected baselines, excelling in color transformation, artistic style transfer, and content preservation. The resulting images exhibit a high degree of visual appeal and fidelity, highlighting the effectiveness of our approach in achieving high-quality style transfer results.

4.3. Quantitative comparison

In addition to the qualitative analysis, we performed a quantitative comparison of our proposed method with the selected baseline methods using Perceptual Similarity Loss [29]. This metric evaluates the perceptual similarity between the generated images and the target content images, providing an objective measure of the effectiveness of the style transfer process.

To compute the Perceptual Similarity Loss, we employed a pre-trained VGG model, which has been widely used in the literature for evaluating style transfer methods. We conducted our quantitative evaluation on a diverse set of over 100 images, ensuring a comprehensive assessment of the performance of our method and the baselines. The results of the quantitative comparison are presented in Tab. 1, which shows the mean Perceptual Similarity Loss values across the 100 images for each method. Our method outperforms the baseline methods, indicating that our approach is more effective in transferring the desired style while preserving the content details of the input image.

	DEPM(ours)	CMD	EFDM	DSTN
LPIPS	**0.596**	0.719	0.606	0.615

Table 1. Quantitative results. Average values of LPIPS Loss (lower is better) are calculated across the 100 images.

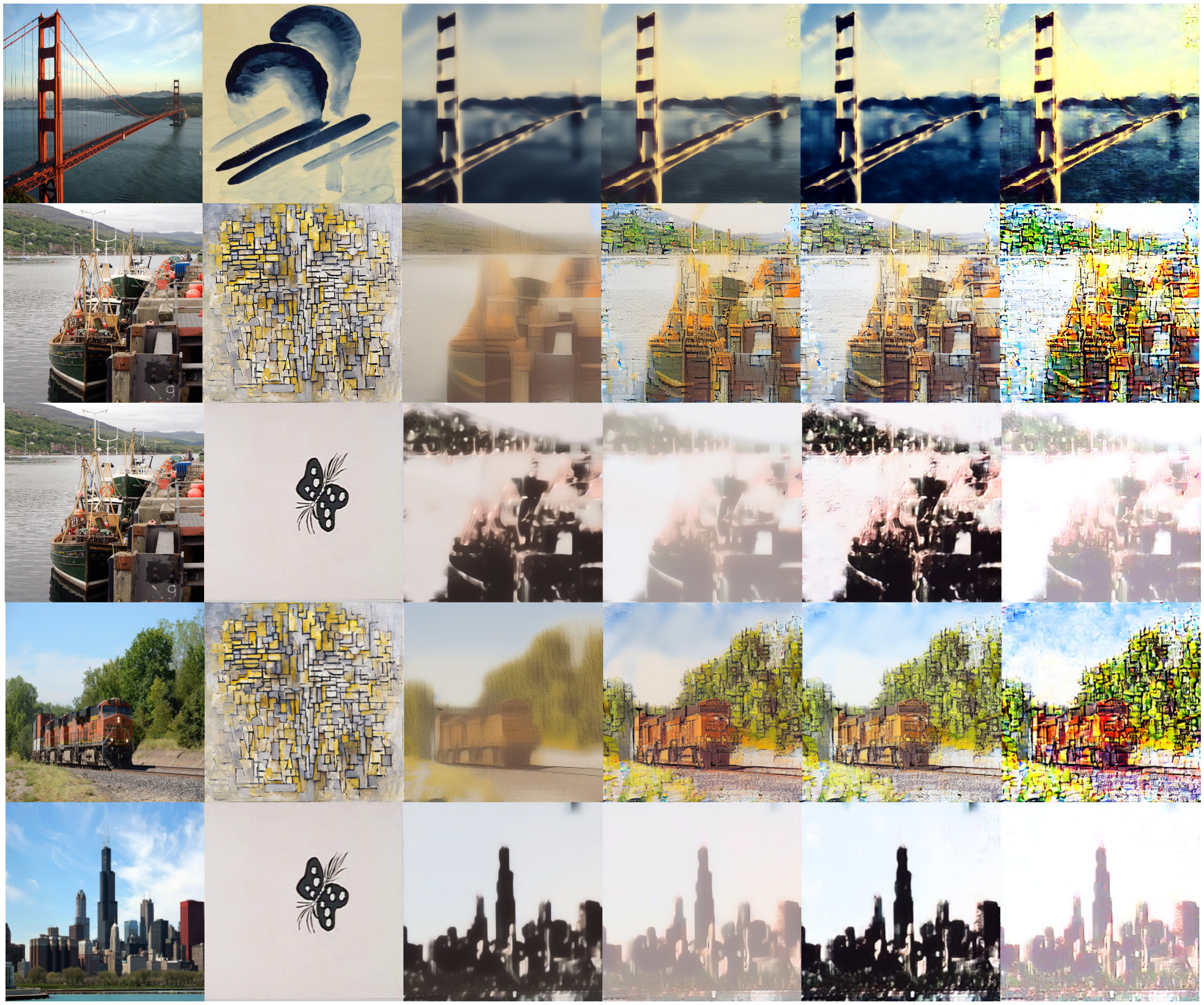
5. Ablation study

In this section, we present an ablation study to investigate the impact of various components of our proposed method on the style transfer results. We conducted extensive experiments by including and excluding the Patch Match (PM), Whitening and Coloring Transform (WCT) whitening and coloring, and constant σ components in different combinations. We also experimented with different versions of the Stable Diffusion model and the application of PM at multiple steps of the denoising process.

The ablation study comprises the following configurations: PM only, PM + WCT, PM * constant σ, and PM * constant σ + WCT, shown in Fig. 4. As discussed in Section 3.2, we implemented the forward diffusion process with $t = 15$ steps of noise to obtain C_{15} and S_{15} and subsequently, during the reverse diffusion, we applied the aforementioned configurations at $t = 15$. Our experiments reveal that the combination of PM multiplied with sigma provides the best balance between artistic style transfer and content preservation. We experimented with Stable Diffusion versions 1.4, 1.5, and 2, and observed no significant differences in the style transfer results across these versions. This indicates that our method is robust to variations in the underlying diffusion model and can be applied to different versions with consistent performance.

Additionally, we explored the application of PM at multiple steps of the denoising process. Fig. 5 Illustrates the visual results of the experiment. Applying PM at multiple steps resulted in increasing deviations of latent codes from Gaussian distribution. A single application of PM near the beginning of the denoising process is optimal for achieving the desired style transfer effects.

Furthermore, we investigated the impact of applying the PM transformation at various steps of the denoising process. Our experiments revealed that the best results were obtained when the PM transformation was applied near the beginning steps, as this allowed for a more accurate and visually appealing transfer of style while preserving the content details



(a) Content (b) Style (c) PM (d) PM + WCT (e) PM $*\sigma$ (f) PM $*\sigma$ + WCT

Figure 4. Comparison of different techniques of DEPM transfers at $t = 15$.

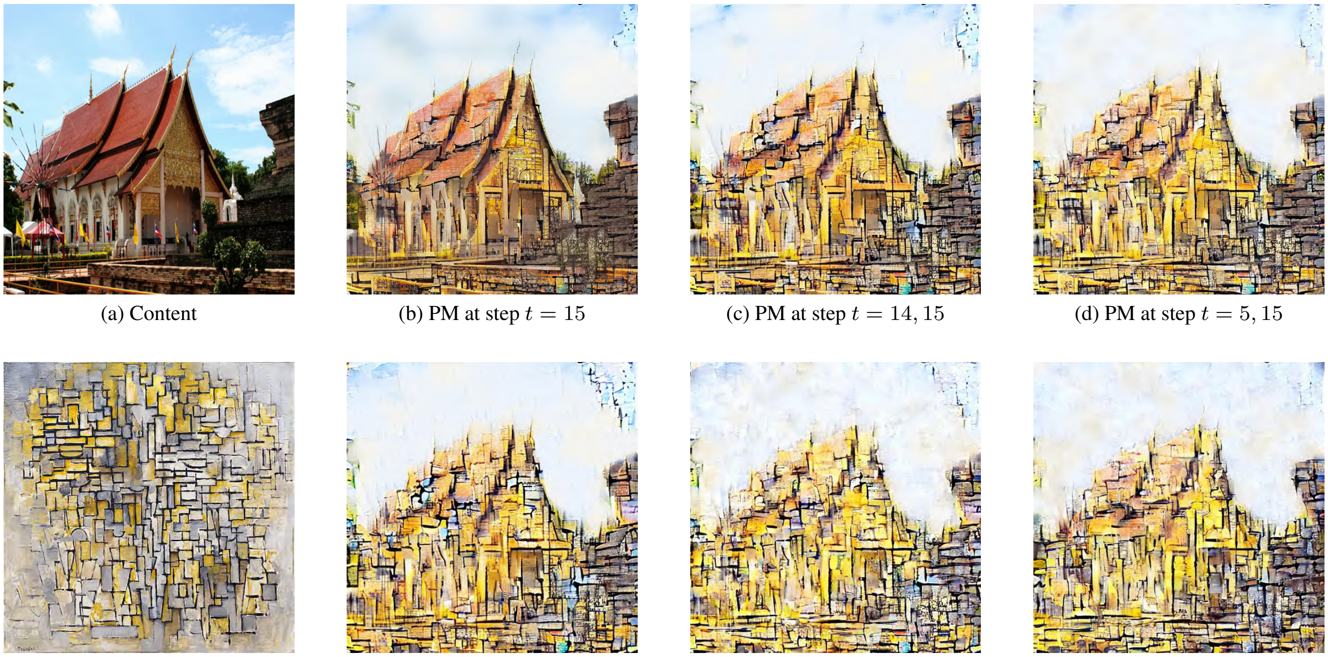
of the input image.

In summary, our ablation study provides valuable insights into the contributions of different components of our proposed method and their impact on the style transfer results. These findings demonstrate the effectiveness of our approach in achieving high-quality style transfer while maintaining the content details of the input image, and highlight the potential for further optimization and refinement of our method.

6. Conclusion

In this paper, we have presented a novel approach to style transfer that synergistically combines diffusion models with style transfer techniques, enabling the transfer of arbitrary styles during the inference step without any finetuning or pretraining. Our method leverages the power of generative diffusion models to capture high-level style features while ensuring that the fine-grained texture details of the original image are preserved.

Through a series of experiments and comparisons with state-of-the-art style transfer methods, we demonstrated the effectiveness of our approach in achieving high-quality style transfer results while preserving the content details of the input image. Our ablation study provided valuable insights into the contributions of different components of our method, such as Patch Match, Whitened Color Transform, and constant sigma, and their impact on the style transfer results.



(a) Content (b) PM at step $t = 15$ (c) PM at step $t = 14, 15$ (d) PM at step $t = 5, 15$

(e) Style (f) PM at step $t = 13, 15, 17$ (g) PM at step $t = 3, 10, 17$ (h) PM at step $t = 3, 5, 7$

Figure 5. Applining PM $*\sigma$ at multiple steps of the denoising process.

Our method offers significant advantages over existing style transfer techniques, such as improved flexibility, efficiency, and the ability to handle arbitrary styles without the need for training. This makes our approach a more accessible and versatile tool for artists, designers, and researchers alike, and encourages further exploration of the potential applications of diffusion models and style transfer techniques.

In conclusion, our work contributes to the growing body of research on diffusion models and style transfer, and opens up new avenues for future research in this area. By building upon the successes of previous works and addressing their limitations, our approach has the potential to advance the state-of-the-art in style transfer research and provide a more effective and efficient solution for a wide range of applications.

References

[1] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style, 2016. 1, 2, 3

[2] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr^2: Image style transfer with transformers, 2021. 2

[3] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023. 1

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 4

[5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 2

[6] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015. cite arxiv:1508.06576. 1, 2

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 1, 2

[8] Kibeom Hong, Seogkyu Jeon, Huan Yang, Jianlong Fu, and Hyeran Byun. Domain-aware universal style transfer, 2021. 2, 5

[9] Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, and Ming-Hsuan Yang. Learning to stylize novel views, 2021. 2

[10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization, 2017. 1

[11] Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis, 2022. 1

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 1, 2

[13] Nikolai Kalischek, Jan Dirk Wegner, and Konrad Schindler. In the light of feature distributions: moment matching for neural style transfer, 2021. 2, 5

[14] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. 4

[15] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 1

[16] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2020. 1

[17] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition, 2021. 2

[18] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation, 2022. 3

[19] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms, 2017. 1, 2, 4

[20] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. 1

[21] Fred Phillips and Brandy Mackintosh. Wiki Art Gallery, Inc.: A Case for Critical Thinking. *Issues in Accounting Education Teaching Notes*, 26(3):53–63, 08 2011. 4

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[24] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. 2

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

[26] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Björn Ommer. A style-aware content loss for real-time hd style transfer, 2018. 1

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2020. 2

[28] Jinchao Yang, Fei Guo, Shuo Chen, Jun Li, and Jian Yang. Industrial style transfer with large-scale geometric warping and content preservation, 2022. 2

[29] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 6

[30] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. 2022. 2, 5