Random Projection-Induced Gaussian Latent Features for Arbitrary Style Transfer

Anonymous authors
Paper under double-blind review

Abstract

The feature transfer technique centered on adjusting mean and variance statistics, widely known as AdaIN, lies at core of current style transfer research. This technique is grounded in the assumption that the latent features for style transfer follow Gaussian distributions. In practice, however, this assumption is often hard to meet, as the features typically exhibit sparse distributions due to the significant spatial correlation inherent in natural images. To tackle this issue, we propose initially projecting the sparse features into lower dimensions via random projection, and then performing style transfer on these projections. Statistically, the projections will follow or approximate Gaussian distributions, thereby better aligning with AdaIN's requirements and enhancing transfer performance. With the stylized projections, we can further reconstruct them back to the original feature space by leveraging compressed sensing theory, thereby obtaining the stylized features. The entire process constitutes a projection-stylization-reconstruction module, which can be seamlessly integrated into AdaIN without necessitating network retraining. Additionally, our proposed module can also be incorporated into the emerging style transfer method based on exchanging cumulative distribution functions, known as EFDM, which faces limitations when there are substantial differences in sparsity levels between content and style features. By projecting both types of features into dense Gaussian distributions, random projection can reduce their sparsity disparity, ultimately improving performance. Experiments demonstrate that the performance improvements mentioned can be achieved on existing state-of-the-art approaches.

1 Introduction

Style transfer aims to transfer the style of one image to another while preserving the semantic content information of the latter. Recent research has shown that this objective can be efficiently realized within a deep encoder-decoder framework, by statistically extracting and integrating style and content information from deep convolutional features (Gatys et al., 2016). In this approach, the statistical modeling of style features plays a pivotal role, yet it remains a challenging issue due to the inherent subjectivity and diversity in defining style features.

There are two fundamental statistical approach to model style features: One approach is grounded in the utilization of mean and variance, famously known as AdaIN (Huang & Belongie, 2017), while the other leverages the cumulative distribution function, termed EFDM (Zhang et al., 2022). Despite their impressive performance, it is noteworthy that both methods have limitations, particularly as deep convolutional features often exhibit sparse distributions (Mahendran & Vedaldi, 2015; Qin et al., 2020), which are not conducive to effective style transfer. The reason is explained as follows. Let us first examine the AdaIN method, which operates on the assumption that deep features have Gaussian distributions, rather than sparse distributions. For a style feature map that exhibits a sparse distribution, its style information is typically concentrated within a few large-magnitude feature elements. However, as the number of small-magnitude feature elements increases, the influence of these large elements on both the mean and variance (essential for AdaIN) tends to decrease. This property suggests that the effect of style transfer with AdaIN may deteriorate, when the style feature map becomes sparser, as demonstrated by the example in Figure 1. In the EFDM method, the cumulative distribution function is adopted to represent style features. The function is transferred by substituting the sorted-feature-elements of the content feature map with corresponding elements from the style feature map. Nevertheless, this method tends to perform worse when the sparsity of the feature distributions is inconsistent between the

Content & Style AdaIN EFDM Ours (AdaIN+RP)

Figure 1: Style transfer using a style image composed of few color points. The methods of AdaIN (Huang & Belongie, 2017) and EFDM (Zhang et al., 2022) lose substantial content details in contrast to our random projection-based style transfer (implemented on AdaIN).

content feature map and the image feature map. This problem is exemplified in Figure 1, where the style feature map is much sparser than the content feature map. It can be seen that the stylized image experiences noticeable content loss, since the content feature map has to lose a few large-magnitude feature elements, after adopting the elements from the *sparser* style feature map. Overall, the aforementioned two problems indicate that the sparse distributions of deep features limit the application effects of AdaIN and EFDM.

To address the challenge, in the paper we propose applying random projection to deep features, before performing style transfer on them. Random projection is a technique that projects high dimensional data to low dimensional subspaces by multiplying the data with a random matrix. Statistically, when the entries of random matrices are drawn from Gaussian distributions, the projections will adhere to Gaussian distribution. Similarly, when the matrix entries are sampled from other sparse distributions, such as $\{0, \pm 1\}$ distributions (Meckes, 2012), the projections will approximate Gaussian distributions by the central limit theorem. These properties suggest that random projection can make features better meet the requirement of AdaIN. Moreover, the dense, Gaussian distribution is favorable to reducing the disparity in sparsity levels between the style image map and the content image map. This, in turn, mitigates the aforementioned sparsity-imbalance issue that can arise in the context of EFDM. By conducting style transfer on the projections of content and style feature maps, we will obtain the stylized projection. To feed the stylized projection into the decoder without network retraining, we further propose to reconstruct the stylized projection back to the original feature map space. This reconstruction relies on compressed sensing theory (Foucart & Rauhut, 2013), which states that a sparse signal can be approximately reconstructed from its random projections, even in the presence of noise. In the context of our research, the stylized projection can be viewed as the projection of content features, perturbed by the projection of style features via style transfer. Then, reconstruction from the stylized projection should yield the desired stylized feature map, which is primarily characterized with content features while being complemented by style features.

The proposed random projection-based style transfer functions as a flexible plug-and-play module, which can be seamlessly integrated into encoder-decoder frameworks for style transfer, without the need for network retraining. Without loss of generality, our research will focus on incorporating this module into two fundamental models: AdaIN (Huang & Belongie, 2017), and EFDM (Zhang et al., 2022). The two models serve as the basis for style transfer and have bee employed in most existing style transfer approaches, including attention mechanism-based models like AdaAttN (Liu et al., 2021) and diffusion model-based approaches such as StyleID (Chung et al., 2024). If performance improvements can be realized with these two basic models, it is reasonable to anticipate similar improvements in other more advanced models that incorporate them. This is validated in our experiments, where our random projection module indeed empowers existing approaches to overcome the limitations in handling sparse features, thereby achieving notable improvements in preserving content details, enriching and diversifying style elements, and elevating the overall perceptual quality.

2 Related Work

Statistical modeling of style features. In the early research of style transfer, Gram matrices are utilized to represent the distribution of style features, achieving impressive results but incurring a relatively high computational burden. To alleviate this issue, AdaIN (Huang & Belongie, 2017) introduces a style modeling approach based on mean and

variance, which offers comparable performance but significantly reduces complexity. Consequently, this approach has found wide application in encoder-decoder-based models (Jing et al., 2020; Chandran et al., 2021; An et al., 2020; Lin et al., 2021; An et al., 2021; Liu et al., 2021), as well as in generative adversarial models (Karras et al., 2019; 2020). To further address the limitation of AdaIN in dealing with sparse features, EFDM (Zhang et al., 2022) has recently proposed modeling style features using cumulative distribution functions. This approach reports more pronounced results than AdaIN, attracting increasing attention in recent research (Kwon et al., 2024; Ge et al., 2024; Zhang et al., 2024b). However, as previously noted, EFDM encounters challenges when content and style feature distributions differ in sparsity levels. In the paper, we will demonstrate that our random projection-based feature sampling module can efficiently tackle the two problems present in AdaIN and EFDM.

Style transfer with global and local features. The primary goal of style transfer is to achieve a desired balance between content and style information. Achieving this goal necessitates a delicate manipulation of both global and local feature transfers. Global features are crucial in maintaining content quality and ensuring style consistency, whereas local features are instrumental for refining local details. Initially, research primarily centers on global, channel-wise deep features, as evidenced in works such as (Gatys et al., 2015a;b; 2016; Dumoulin et al., 2016; Johnson et al., 2016; Ulyanov et al., 2016; Gatys et al., 2017; Risser et al., 2017; Huang & Belongie, 2017; Li et al., 2017; 2019; 2020; Huang & Gupta, 2020; Zhang et al., 2022). Subsequently, the focus shifts towards the modeling and transferring on local patches or patterns, as demonstrated in (Chen & Schmidt, 2016; Gu et al., 2018; Zhang et al., 2019; Wang et al., 2021; Hong et al., 2023; Park & Lee, 2019). The transition has driven the emergence of attention-based approaches (Chen et al., 2021; Liu et al., 2021; Wu et al., 2021; Deng et al., 2022; Huang et al., 2023; Xu et al., 2023; Zhu et al., 2023; Ma et al., 2023; Zhang et al., 2024a). Recently, diffusion model-based approaches have significantly boosted style manipulation capabilities. Specifically, methods such as Prompt-to-Prompt (Hertz et al., 2022) and Plug-and-Play (Tumanyan et al., 2023) utilize cross-attention and self-attention maps to guide spatial arrangements during text-driven image editing. More recently, StyleID (Chung et al., 2024) has pushed this paradigm further by integrating key and value features from style images into self-attention layers, facilitating training-free artistic style transfer with precise spatial and statistical alignment. Despite these advancements, efficiently modeling and balancing between global and local features remain challenging, due to the elusive nature of style definition and perception, as well as the complexity of style transfer. In the paper, we achieve a flexible balance between global and local feature transfers, by leveraging the local patch sampling with random projection and the feature grouping on projections.

3 Methodology

As discussed before, the proposed random projection-based style transfer can be incorporated into arbitrary encoder-decoder framework, without the need for network retraining. In this section, we elaborate this incorporation process with the basic approach AdaIN, as illustrated in Figure 2, and the incorporation into other approaches, like EFDM, StyleID and AdaAttN, can be realized in the similar way, as outlined in the appendix.

In Figure 2, we provide the overview of incorporating the random projection module into AdaIN, which mainly consists of three steps. First, the feature maps of the content and style images are projected to low-dimensional spaces by random projection. Then style transfer is conducted over the projections of the two kinds of features. To achieve a balance between local details and global consistency, we here divide each projection into a number of subgroups and conduct style transfer on each subgroup. Finally, with the stylized projection, the target image feature map is derived by sparse reconstruction. Overall, the above three steps constitute a projection-stylization-reconstruction module, which is detailed as follows.

3.1 Random projection

Prior to introducing random projection, let us first review the feature feed-forward process involved in the encoder-decoder model, as illustrated in Figure 2. The model takes as inputs a content image I_c and an arbitrary style image I_s, and outputs a target image I_t that combines the semantic content of I_c and the style of I_s. This combination is performed in a single or a few layers of the encoder. In each layer, we can derive two feature maps respectively for I_c and I_s. Usually, each feature map will consist of multiple channels. For I_c, we denote the feature map in each channel with a vector $f_c \in \mathbb{R}^{(H_c \times W_c) \times 1}$, where H_c and W_c indicate the height and width of the content feature map. In the similar way, we can define $f_s \in \mathbb{R}^{(H_s \times W_s) \times 1}$ for I_s. Note that we here describe the feature of each channel, rather

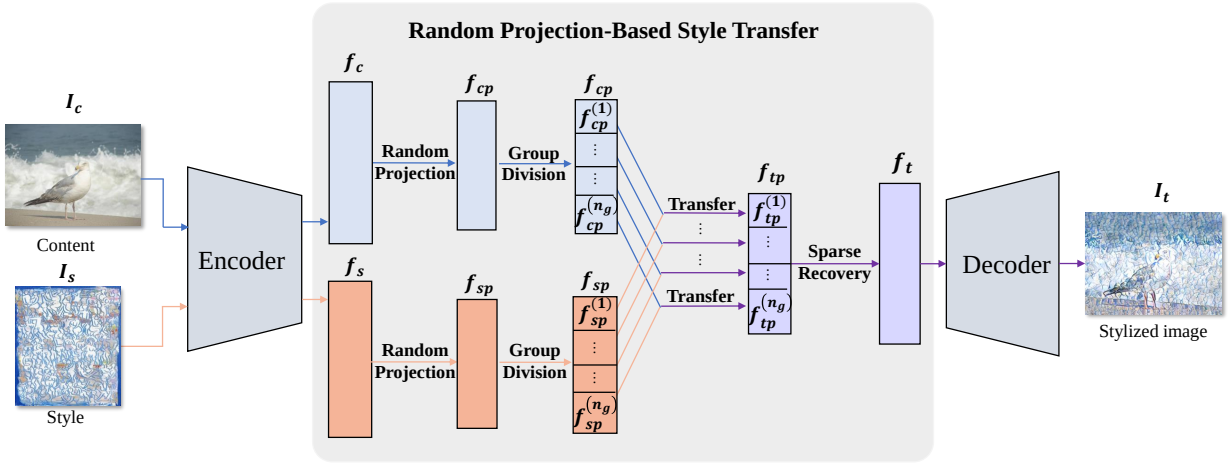

Figure 2: Overview of the random projection-based style transfer implemented on AdaIN. By feeding the content image I_c and the style image I_s into the encoder, we derive the content feature map f_c and the style feature map f_s, which are further projected to low dimensional spaces f_{cp} and f_{sp} by random projection. The projections are divided into a number n_g of subgroups: $f_{cp} = [f_{cp}^{(1)}, \cdots, f_{cp}^{(n_g)}]$ and $f_{sp} = [f_{sp}^{(1)}, \cdots, f_{sp}^{(n_g)}]$; and style transfer is conducted on each pair of subgroups $f_{cp}^{(i)}$ and $f_{sp}^{(i)}$, resulting in the stylized projection $f_{tp} = [f_{tp}^{(1)}, \cdots, f_{tp}^{(n_g)}]$. With the projection f_{tp}, by sparse reconstruction we derive its counterpart in the original feature space, namely the target feature map f_t . Feeding f_t into the decoder, the desired target (stylized) image I_t is finally obtained.

than the features of all channels, since the random projection-based style transfer is conducted on channel-wise basis. Given the content feature map f_c and the style feature map f_s, we then can derive their random projections

$$f_{cp} = A_c f_c \quad \text{and} \quad f_{sp} = A_s f_s \tag{1}$$

where $A_c \in \mathbb{R}^{M_c \times (H_c \times W_c)}$ and $A_s \in \mathbb{R}^{M_s \times (H_s \times W_s)}$ are the random projection matrices.

For random projection, the choice of random matrices is crucial. In this paper, we investigate four commonly-used random matrices in compressed sensing: the {0,1}-matrix with $\sqrt{M_c}$ nonzero elements per column (Lu et al., 2018), the {1,-1}- and {0,1,-1}-matrices with elements taken with equal probabilities (Achlioptas, 2003; Amini & Marvasti, 2011), and the Gaussian matrix (Candes & Tao, 2005). It can be seen that the {0,1}-matrix exhibits very sparse structures compared to other three kinds of matrices. Empirically, the three relatively dense matrices perform well when the style feature map exhibits very sparse distributions. This is because the random projection based on dense matrices is more likely to yield Gaussian-like distributions. In contrast, when the style feature is not very sparse, the sparse {0,1}-matrix tends to perform better. The advantage may be explained with the following fact. Essentially, the {0,1}-matrix based random projection performs a random, sparse sampling over the features f_c and f_s. The resulting projections $f_{cp} \in \mathbb{R}^{M_c}$ and $f_{sp} \in \mathbb{R}^{M_s}$ have their each element related to a few elements of the original features f_c and f_s, reflecting the local correlation of original features. This allows the stylization on feature projections to better capture the local details compared to stylization on original features.

In addition to the distribution of random matrices, style transfer is also related to the compression ratio of random matrices, which can be written as $r = M_c/(H_c \times W_c)$. With the decreasing of compression rate r, the following sparse reconstruction will become hard and then introduce noise. Empirically, the increased noise tends to degrade the quality of content, but enrich the diversity of style. In addition, the random variation of the random matrices themselves can also result in diverse styles.

3.2 Grouping-based style transfer

With the feature projections $f_{cp} \in \mathbb{R}^{M_c}$ and $f_{sp} \in \mathbb{R}^{M_s}$ derived in the previous random projection, we are ready to perform style transfer on them. To control the impact of style transfer on local details, we propose to first divide the feature projections into a number n_g of subgroups, namely having $f_{cp} = [f_{cp}^{(1)}, \cdots, f_{cp}^{(n_g)}]$ and $f_{sp} = [f_{sp}^{(1)}, \cdots, f_{sp}^{(n_g)}]$,

and then perform style transfer on each pair of subgroups $f_{cp}^{(i)} \in \mathbb{R}^{\frac{M_c}{n_g}}$ and $f_{sp}^{(i)} \in \mathbb{R}^{\frac{M_s}{n_g}}$. It is evident that the greater the number n_g of subgroups, the fewer local elements are involved in style transfer. By adjusting the value pf n_g, we can strike a balance between local details and global consistency.

By conducting AdaIN on each pair of subgroups $f_{cp}^{(i)}$ and $f_{sp}^{(i)}$, the stylized feature is derived as

$$f_{tp}^{(i)} = \sigma(f_{sp}^{(i)}) \frac{f_{cp}^{(i)} - \mu(f_{cp}^{(i)})}{\sigma(f_{cp}^{(i)})} + \mu(f_{sp}^{(i)}). \tag{2}$$

Combing the results $f_{tp}^{(i)} \in \mathbb{R}^{\frac{M_c}{n_g}}$ of all subgroups, we can obtain the stylized feature on the entire projection $f_{tp} = [f_{tp}^{(1)}, \cdots, f_{tp}^{(n_g)}] \in \mathbb{R}^{M_c}$.

Similarly as the content projection f_{cp} and the style projection f_{sp} as derived in equation 1, we here hypothesize that the stylized projection f_{tp} is a random projection of an underlying target (stylized) feature map f_t, namely

$$f_{tp} = A_c f_t. \tag{3}$$

Here, we exploit the same projection matrix A_c with the content projection f_{cp}. This choice is based on the fact that the stylized projection f_{tp} is predominantly characterized with the content projection f_{cp}, and is slightly altered by the style projection f_{sp} via style transfer. According to compressed sensing theory (Foucart & Rauhut, 2013), a target feature map f_t similar to the content feature map f_c can be reconstructed through sparse reconstruction, if the noise introduced to the content projection f_{cp}, namely the variation induced by style transfer, is limited. In general, the variation in f_{cp} during style transfer is more significant when using EFDM compared to AdaIN. The difference arises because AdaIN changes only the mean and variance of f_{cp}, whereas EFDM modifies the cumulative distribution function by replacing all elements of f_{cp} with those of f_{tp}.

3.3 Sparse Reconstruction

With the projection hypothesis equation 3, we now need to reconstruct the target image feature map f_t from its projection f_{tp}. By compressed sensing, f_t can be approximately derived by

$$\hat{f}_t = \arg\min_{f_t} \|A_c f_t - f_{tp}\|_2^2 + \lambda \left|f_t\right|_1, \tag{4}$$

if the feature f_t has adequately sparse distributions, and the projection matrix A_c has sufficiently low correlations between columns. The two conditions should be approximately met, as the deep convolutional feature f_t is usually sparse, and the random matrices employed here are commonly utilized in compressed sensing. After obtaining the target feature map f_t, we can further feed it into the encoder-decoder model to generate the desired target image I_t.

As a convex problem, equation 4 can be tackled with standard optimization algorithms. Considering deep convolution features usually have high dimensions, it is necessary to select an efficient algorithm. In our experiments, we adopt the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) (Beck & Teboulle, 2009), which exhibits a complexity of $\mathcal{O}(MN)$ per iteration, given a projection matrix of size $M \times N$. This algorithm supports GPU-accelerated parallel computation (Feinman, 2021), achieving rapid performance. For instance, it takes about 0.1 seconds to reconstruct a 512×4096-sized feature from its 512×2048-sized projection when running on a GeForce RTX 3080 GPU.

4 Experiments

Based on the analysis presented in the previous section, our random projection-based style transfer method is subject to the influence of several parameters tied to random projection. These include the distribution and compression rate r of random matrices, as well as the number n_g of subgroups into which the feature projections f_{cp} and f_{sp} are divided. In this section, we will first examine how these parameters impact stylized results, enabling us to achieve a desired balance between style and content by reasonably adjusting the parameters. Then, we will showcase the performance improvements achieved by integrating our approach into two fundamental style transfer models: AdaIN (Huang & Belongie, 2017) and EFDM (Zhang et al., 2022), as well as into their advanced variants and other state-of-the-art approaches, such as the diffusion model-based StyleID (Chung et al., 2024) and the attention mechanism-based AdaAttN (Liu et al., 2021). Prior to these studies, we will first outline the implementation details.

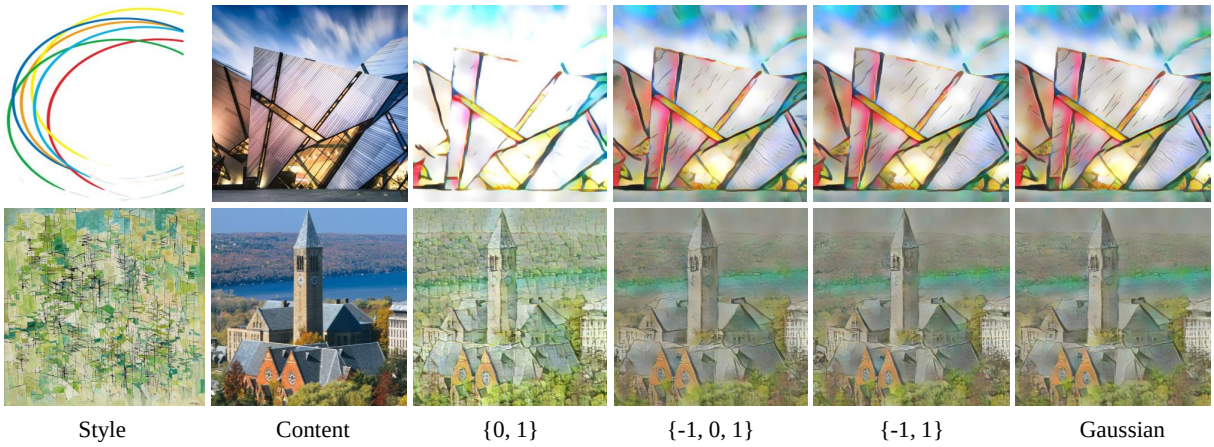

Figure 3: Stylized results derived with four different random matrices: {0, 1}, {-1, 0, 1}, {-1, 1} and Gaussian matrices.

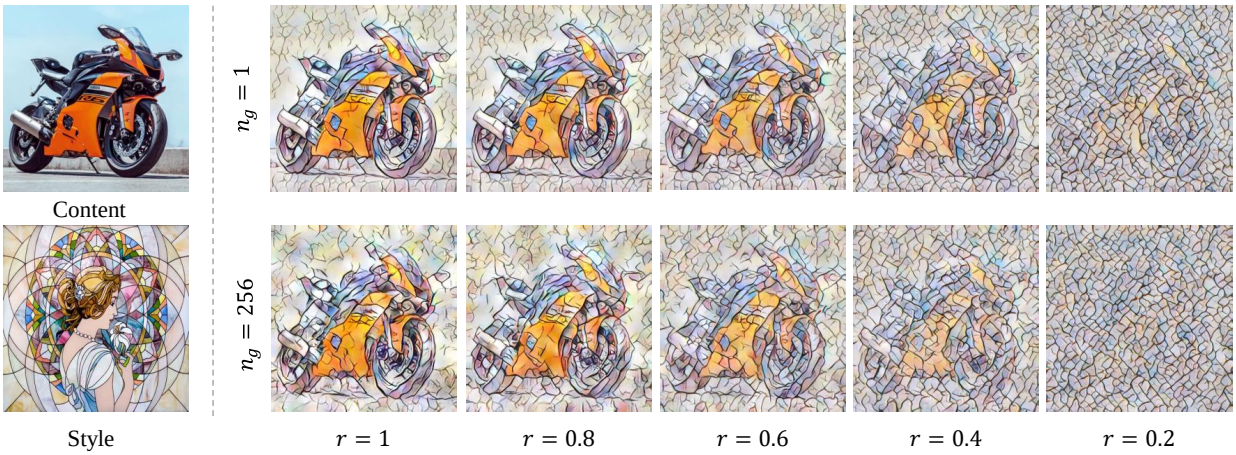

Figure 4: Stylized results derived by varying the compression rate r of random matrices from 0.2 to 1, and the number of subgroups n_g within feature projections from 1 to 256. Random projection is implemented with {0, 1} matrices.

4.1 Implementation Details

Given the models AdaIN, AdaAttN, EFDM and StyleID, as illustrated in Figure 2, our random projection-based style transfer module is incorporated into each layer of these models that involve style transfer, without the need for network retraining. The computation cost introduced by our module primarily stems from the sparse reconstruction outlined in equation 4, which is implemented using the FISTA algorithm (Beck & Teboulle, 2009; Feinman, 2021). For the algorithm, we assign the regularization parameter of $\lambda = 0.5$ and limit the maximum number of iterations to 10. Regarding random projection, we set the compression ratio to $M/N = 0.8$. For style transfer, we set the number of subgroups within feature projections to $n_g = 1$. Overall, our performance advantage is not overly sensitive to variations in these parameters. The code will be made publicly available.

4.2 Impact of the parameters of random projection on style transfer

Distribution of random matrices. In Figure 3, we compare the stylized results derived with four popular random projection matrices: {0, 1}, {-1, 0, 1}, {-1, 1}, and Gaussian matrices. It is evident that when the style feature exhibits a sparse distribution, as illustrated in the first row, the {0, 1}-matrix notably underperforms compared to the other three matrices, while the latter three exhibit comparable performance. Notably, the {0, 1}-matrix comprises only a ratio of

$1/\sqrt{M_c}$ nonzero entries, leading to much sparser distributions than the other matrices. Consequently, it struggles in projecting the style feature towards denser, Gaussian distributions, resulting in inferior style transfer performance. Conversely, the $\{0, 1\}$-matrix excels when the style feature displays relatively dense distributions, as depicted in the second row of Figure 3. In such scenarios, generating projections towards Gaussian distributions becomes more feasible. In our experiments, unless otherwise specified, we will generally use Gaussian matrices.

Compression rate of random matrices and the number of subgroups within feature projections. In Figure 4, we compare the stylized results by varying the compression rate r of random $\{0, 1\}$-matrices from 0.2 to 1, and the number n_g of subgroups within feature projections from 1 to 256. It is apparent that as the compression rate r decreases or the subgroup number n_g increases, the style gradually becomes prominent while the content becomes less evident. The impact of compression rates can be understood as follows. In compressed sensing, a lower compression rate r will lead to a higher reconstruction noise in the reconstructed stylized feature f_t. Then the increased noise introduces more irregular style patterns, while degrading the content. Regarding the increase in the number of subgroups n_g, as mentioned in the previous section, it implies that the number of feature elements within each subgroup for style transfer decreases. This element-restricted style transfer may overly focus on local details and neglect global constraints, leading to content loss. Throughout the experiments, we generally set the compression rate $r = 0.8$ and the number of subgroups $n_g = 1$.

4.3 Improvement of our random projection module over state-of-the-art approaches

Improvement over two fundamental models: AdaIN and EFDM. In Figure 5, we stylize four different content images using four distinct style images with relatively sparse features. Without sacrificing generality, the content images encompass four distinct attribute categories: human figures, animals, plants, and architectural structures. Meanwhile, the style images display two different types of sparse distributions: linear patterns and patchy textures. It is clear that both AdaIN and EFDM result in a considerable loss of texture, color, and content details, leading to poor perceptual quality. As previously discussed, this limitation arises because these two models are not well suited for handling style features with sparse distributions. As expected, when we incorporating our random projection module into the two models, the resulting AdaIN+RP and EFDM+RP model achieve substantial performance improvements by transforming the features from sparse distributions toward Gaussian distributions. Our success with the two foundational models not only validates the effectiveness of our method but also suggests its broad applicability, given that the two models form the core of most existing style transfer approaches, as elaborated later.

Improvement over the state-of-the-art, diffusion model-based approach: StyleID. To validate the effectiveness and generalizability of our method, we examine how it improves the performance of the state-of-the-art approach StyleID by keeping either the content or the style constant while varying the other, as depicted in Figures 6 and 7. In Figure 6, a facial image undergoes stylization with various style images. It is evident that across the various styles, our method consistently brings about significant enhancements in texture details and contrast, color saturation, and overall visual perception. Similar improvements are also noticeable in Figure 7, where various content images are stylized using a black-lined portrait sketch featuring red lips. Particularly, it can be seen that our method successfully and significantly transfers the color of the red lips, even though this color constitutes only a very small portion of the entire style image. This validates the advantage of our random projection module in capturing minor but crucial sparse features.

Improvement over the state-of-the-art, attention-based approach: AdaAttN. Figure 8 illustrates notable performance improvements in texture clarity and contrast, when integrating our random projection module into the AdaAttN. Furthermore, the resulting AdaAttN+RP also outperforms other well-recognized approaches, including AesPA(Hong et al., 2023), RAST(Ma et al., 2023), StyTr^2(Deng et al., 2022), IEST(Chen et al., 2021), Styleformer(Wu et al., 2021), and Artflow(An et al., 2021), as demonstrated in the appendix, specifically in Figures 11 to 13. Note that our AdaAttN+RP excels not only with visually sparse-style images, such as the ones depicted in Figures 11, but also with those that are visually dense, exemplified by Figures 13. This wider applicability stems from the fact that the visually dense images also tend to exhibit sparse feature distributions because of the high spatial correlation between image pixels. As a result, our random projection module can be used to further enhance the style transfer performance by transforming these sparse feature distributions towards the desired Gaussian distributions.

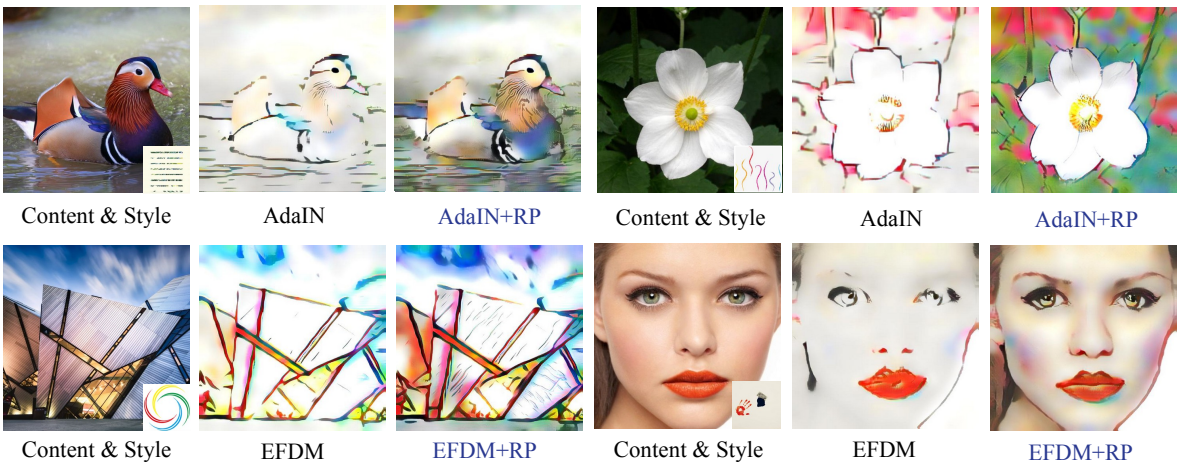

Figure 5: Stylized results derived by incorporating vs. *not* incorporating our random projection (RP) module into the two basic models: AdaIN and EFDM, respectively in the upper and lower rows.

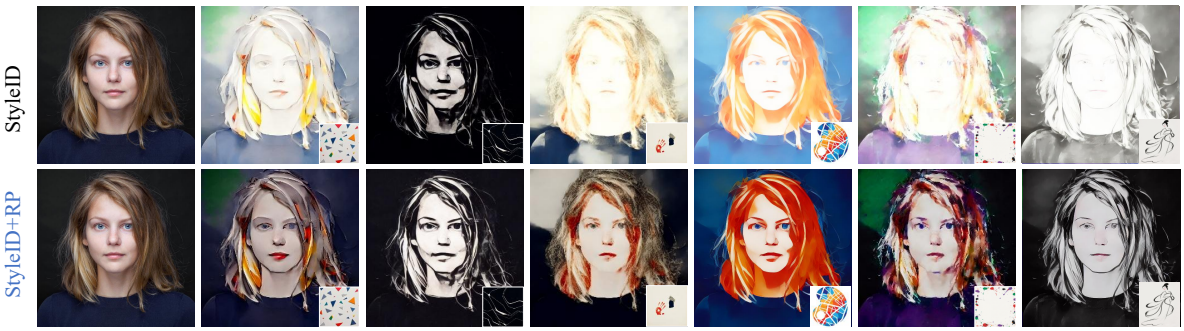

Figure 6: Stylized results derived by incorporating vs. *not* incorporating our random projection (RP) module into the state-of-the-art, diffusion model-based approach: StyleID, with a fixed content but various styles.

Figure 7: Stylized results derived by incorporating vs. *not* incorporating our random projection (RP) module into the state-of-the-art, diffusion model-based approach: StyleID, with various contents but a fixed style (a black-line portrait with red lips).

Figure 8: Stylized results derived by incorporating vs. *not* incorporating our random projection (RP) module into AdaAttN.

Table 1: Quantitative results of LPIPS, ArtFID and User Study. The better results are marked with an underline.

Metric	AdaIN	AdaIN+RP	EFDM	EFDM+RP	StyleID	StyleID+RP
LPIPS ↓	0.665	<u>0.579</u>	0.686	<u>0.592</u>	0.650	<u>0.568</u>
ArtFID ↓	31.01	<u>29.65</u>	33.88	<u>31.37</u>	28.51	<u>26.42</u>
Preference ratio↑	0.26	<u>0.74</u>	0.19	<u>0.81</u>	0.13	<u>0.87</u>

Quantitative results. The results shown in Figures 5-8 demonstrate that our random projection module significantly improves the quality of content, delivering superior perceptual performance in style transfer. This property is further supported by the quantitative results presented in Table 1, where LPIPS (Zhang et al., 2018) measures the content fidelity of stylized results, and ArtFID (Wright & Ommer, 2022) assesses the overall style transfer performance by considering both content and style qualities. Moreover, we conducted a user study with 50 participants who selected their preferred stylized results from 60 trials based on overall perceptual quality. All these results validate the advantage of our approach.

5 Conclusion

Existing style transfer approaches based on the encoder-decoder architecture typically employ either the mean-and-variance strategy (proposed in AdaIN) or the cumulative distribution function (introduced in EFDM) to model the distribution of style features. The two methods implicitly favor the scenarios where the features of both style and content images roughly follow Gaussian distributions. Nevertheless, this requirement is often hard to satisfy, as deep encoder features usually exhibit sparse distributions. Consequently, when handling highly sparse features, the two methods often result in poor perceptual quality, with a significant loss of texture details, contrast, color, and content. In this paper, we have for the first time identified and tackled this issue by incorporating a random projection module into the encoder-decoder architecture, in order to transform features from sparse distributions towards the desired Gaussian distributions. This module operates as a flexible plug-and-play component, without the need for network retraining. When integrated into the fundamental models AdaIN and EFDM, as well as other cutting-edge approaches like the diffusion model-based StyleID and the attention mechanism-based AdaAttN, this module brings about notable performance improvements, outperforming other well-recognized approaches.

References

D. Achlioptas. Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

Arash Amini and Farokh Marvasti. Deterministic construction of binary, bipolar, and ternary compressed sensing matrices. *IEEE Transactions on Information Theory*, 57(4):2360–2370, 2011.

Jie An, Tao Li, Haozhi Huang, Li Shen, Xuan Wang, Yongyi Tang, Jinwen Ma, Wei Liu, and Jiebo Luo. Real-time universal style transfer on high-resolution images via zero-channel pruning. *arXiv preprint arXiv:2006.09029*, 2020.

Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 862–871, 2021.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203 – 4215, Dec. 2005.

Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, Markus Gross, and Derek Bradley. Adaptive convolutions for structure-aware style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7972–7981, 2021.

Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34:26561–26573, 2021.

Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8795–8805, 2024.

Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. StyTr^2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11326–11336, 2022.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.

Reuben Feinman. A pytorch library for L1-regularized least squares (lasso) problems. 2021.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013. ISBN 0817649476.

Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015a.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015b.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3985–3993, 2017.

Bin Ge, Zhenshan Hu, Chenxing Xia, and Junming Guan. Arbitrary style transfer method with attentional feature distribution matching. *Multimedia Systems*, 30(2):96, 2024.

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.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Kibeom Hong, Seogkyu Jeon, Junsoo Lee, Namhyuk Ahn, Kunhee Kim, Pilhyeon Lee, Daesik Kim, Youngjung Uh, and Hyeran Byun. Aespa-net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22758–22767, 2023.

Eddie Huang and Sahil Gupta. Style is a distribution of features. *arXiv preprint arXiv:2007.13010*, 2020.

Siyu Huang, Jie An, Donglai Wei, Jiebo Luo, and Hanspeter Pfister. Quantart: Quantizing image style transfer towards high visual fidelity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5947–5956, 2023.

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.

Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. Dynamic instance normalization for arbitrary style transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 4369–4376, 2020.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Joonwoo Kwon, Sooyoung Kim, Yuewei Lin, Shinjae Yoo, and Jiook Cha. Aesfa: An aesthetic feature-aware arbitrary neural style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13310–13319, 2024.

Pan Li, Dan Zhang, Lei Zhao, Duanqing Xu, and Dongming Lu. Style permutation for diversified arbitrary style transfer. *IEEE Access*, 8:199147–199158, 2020.

Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3809–3817, 2019.

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.

Tianwei Lin, Zhuoqi Ma, Fu Li, Dongliang He, Xin Li, Errui Ding, Nannan Wang, Jie Li, and Xinbo Gao. Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5141–5150, 2021.

Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6649–6658, 2021.

W. Lu, T. Dai, and S. T. Xia. Binary matrices for compressed sensing. *IEEE Transactions on Signal Processing*, 66 (1):77–85, Jan. 2018. ISSN 1053-587X.

Ariel Lubonja, Sebastian Kazmarek Præsius, and Trac Duy Tran. Efficient batched cpu/gpu implementation of orthogonal matching pursuit for python, 2024.

Yingnan Ma, Chenqiu Zhao, Xudong Li, and Anup Basu. Rast: Restorable arbitrary style transfer via multi-restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 331–340, 2023.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.

Elizabeth Meckes. Approximation of projections of random vectors. *Journal of Theoretical Probability*, 25:333–352, 2012.

Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5880–5888, 2019.

Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, and Nicu Sebe. Binary neural networks: A survey. *Pattern Recognition*, 105:107281, 2020.

Eric Risser, Pierre Wilmot, and Connelly Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, 2017.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023.

Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv preprint arXiv:1603.03417*, 2016.

Zhizhong Wang, Lei Zhao, Haibo Chen, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Divswapper: Towards diversified patch-based arbitrary style transfer. *arXiv preprint arXiv:2101.06381*, 2021.

Matthias Wright and Björn Ommer. Artfid: Quantitative evaluation of neural style transfer. In *DAGM German Conference on Pattern Recognition*, pp. 560–576. Springer, 2022.

Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14618–14627, 2021.

Wenju Xu, Chengjiang Long, and Yongwei Nie. Learning dynamic style kernels for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10083–10092, 2023.

Chiyu Zhang, Xiaogang Xu, Lei Wang, Zaiyan Dai, and Jun Yang. S2wat: Image style transfer via hierarchical vision transformer using strips window attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7024–7032, 2024a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Tianlong Zhang, Jing Lv, and Ming Yang. Semi-parametric style transfer with multi-perspective feature fusion and information-guided alignment. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 943–950, 2024b.

Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8035–8045, 2022.

Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, and Jimei Yang. Multimodal style transfer via graph cuts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5943–5951, 2019.

Mingrui Zhu, Xiao He, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. All-to-key attention for arbitrary style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23109–23119, 2023.

A Appendix

A.1 Overview of incorporating random projection into StyleID

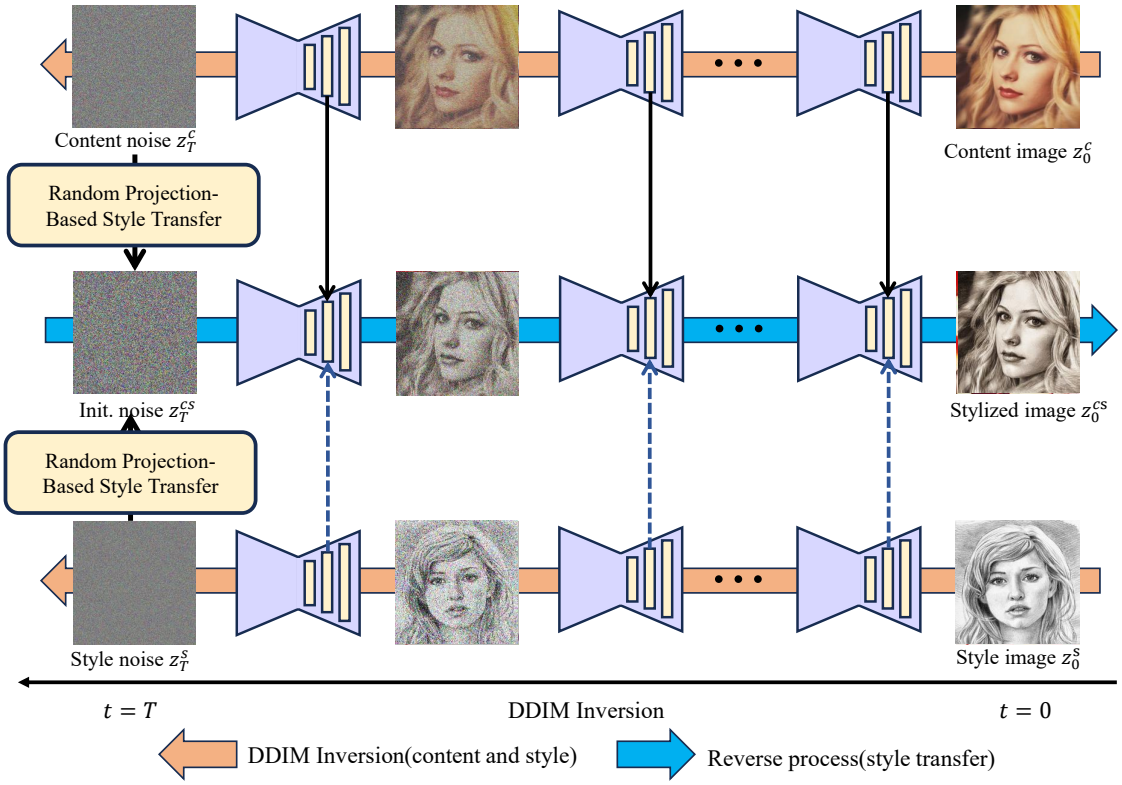

Figure 9: Overview of the random projection-based style transfer implemented on StyleID. First, content and style images undergo identical DDIM inversion process, with attention-based style injection module strictly following StyleID's original implementation. Subsequently, our random projection-based style transfer module operates within StyleID's Initial Latent AdaIN module, performing the operations of random projection and style transfer as illustrated in Figure 2. Finally, the modified latent progresses through StyleID's reverse diffusion process for image synthesis.

A.2 Overview of incorporating random projection into AdaAttN

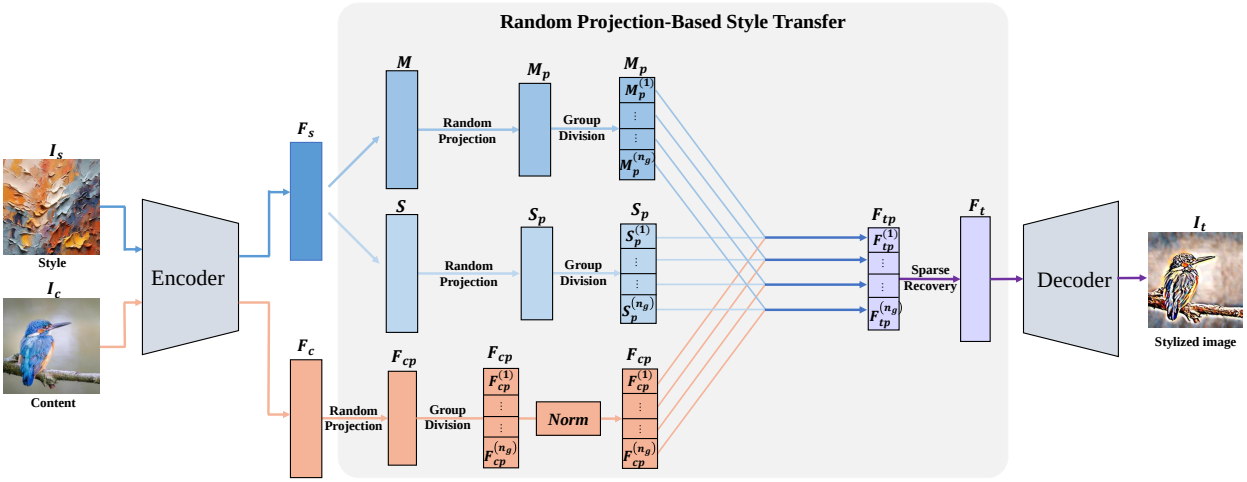

Figure 10: Overview of the random projection-based style transfer implemented on AdaAttN. By feeding the style image I_s and the content image I_c into the VGG encoder, we derive the style feature map F_s and the content feature map F_c. Then the weighted mean M and standard variance S are obtained from F_s, and the specific method is mentioned in AdaAttN (Liu et al., 2021). M, S and F_c are further projected to low dimensional spaces M_p, S_p and F_{cp} by random projection. The projections are divided into a number n_g of subgroups: $M_p = [M_p^{(1)}, \cdots, M_p^{(n_g)}]$, $S_p = [S_p^{(1)}, \cdots, S_p^{(n_g)}]$ and $F_{cp} = [F_{cp}^{(1)}, \cdots, F_{cp}^{(n_g)}]$, and F_{cp} is further mean-variance channel-wise normalized, represented by $Norm$ in the figure. Then style transfer is conducted on each pair of subgroups $M_p^{(i)}$, $S_p^{(i)}$ and $F_{cp}^{(i)}$, resulting in the stylized projection $F_{tp} = [F_{tp}^{(1)}, \cdots, F_{tp}^{(n_g)}]$. The formula for style transfer is outlined as $F_{tp}^{(i)} = F_{cp}^{(i)} * S_p^{(i)} + M_p^{(i)}$. With the projection F_{tp}, by sparse reconstruction we derive its counterpart in the original feature space, namely the target feature map F_t. Feeding F_t into the decoder, the desired target (stylized) image I_t is finally obtained.

A.3 Comparison between our AdaAttN+RP and other well-recognized approaches

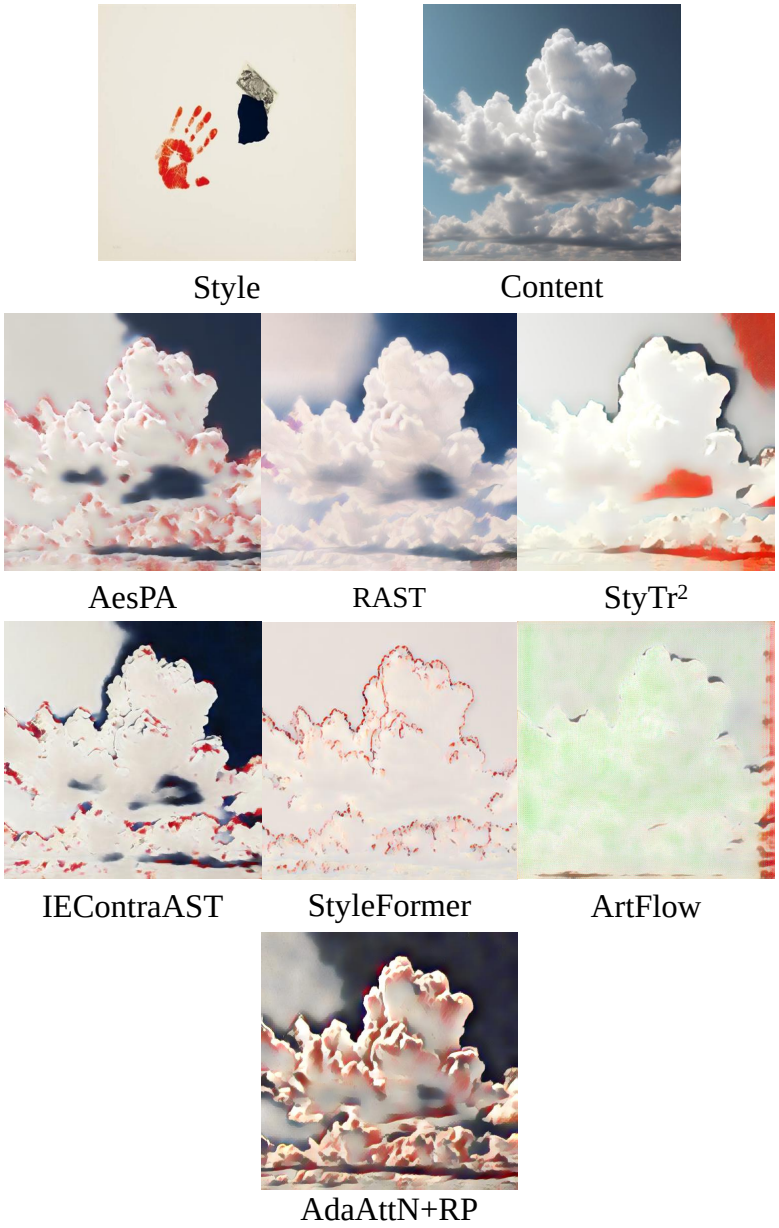

Figure 11: Stylized results derived by our AdaAttN+RP and other approaches. Among the approaches, our AdaAttN+RP stands out for its more saturated and evenly distributed red color.

Table 2: User Study Results for Figure 11

	AesPA	RAST	StyTr^2	IEContraAST	StyleFormer	ArtFlow	AdaAttN+RP
Preference ratio ↑	0.13	0.00	0.00	0.07	0.02	0.00	0.78

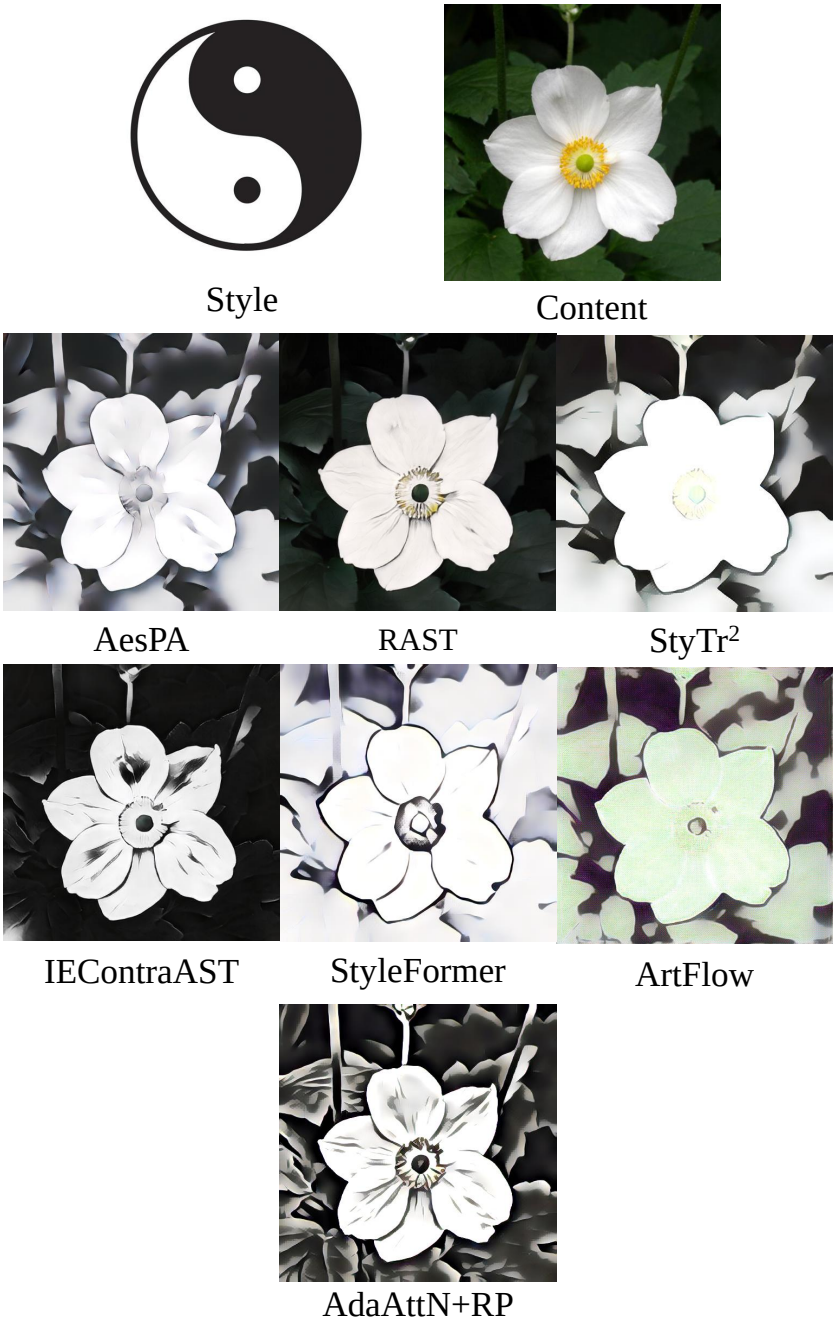

Figure 12: Stylized results derived by our AdaAttN+RP and other approaches. When rendering the textures of branches, leaves, and petals in a black-and-white style, our AdaAttN+RP outperforms all others.

Table 3: User Study Results for Figure 12

	AesPA	RAST	StyTr^2	IEContraAST	StyleFormer	ArtFlow	AdaAttN+RP
Preference ratio ↑	0.08	0.13	0.00	0.01	0.05	0.00	0.73

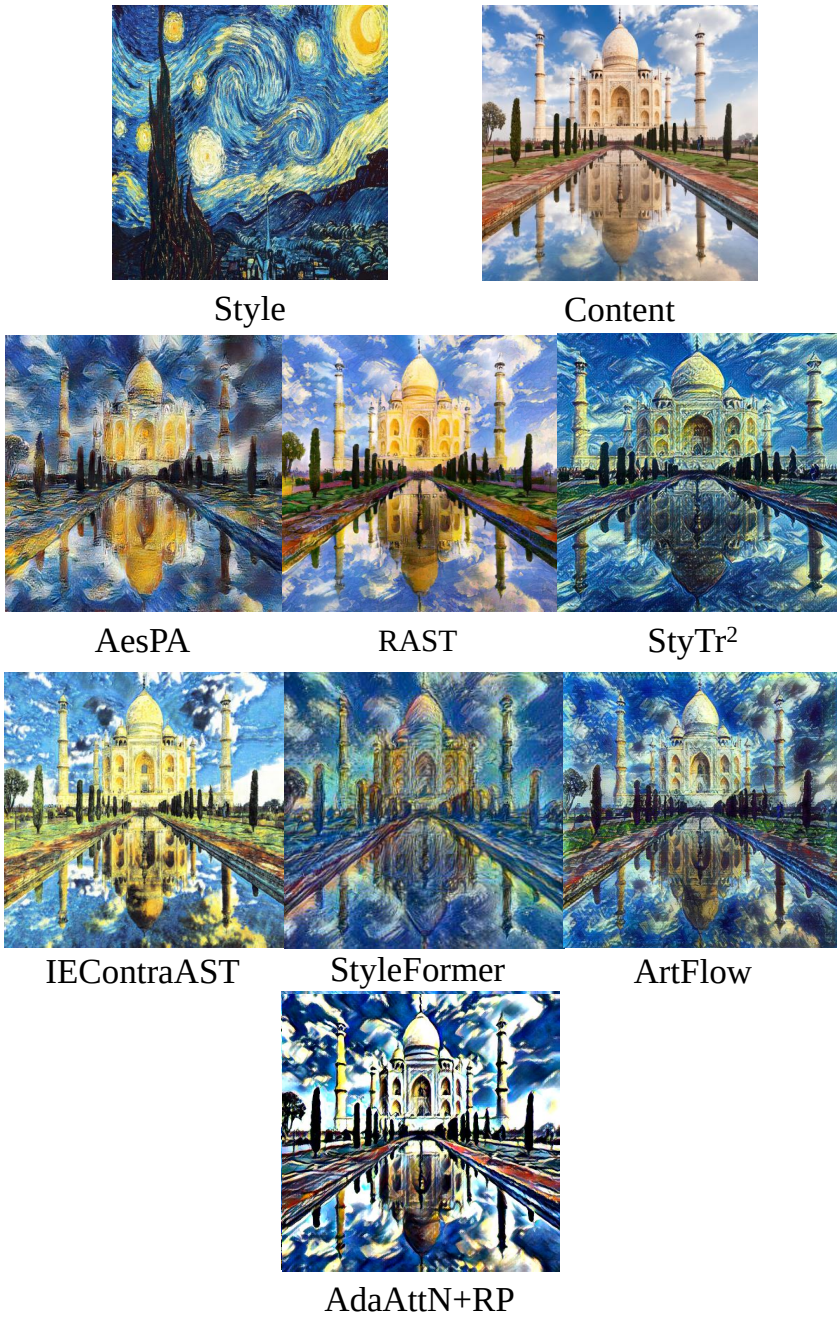

Figure 13: Stylized results derived by our AdaAttN+RP and other approaches. Our AdaAttN+RP achieves a good balance between style transfer and content preservation. In contrast, other approaches either overemphasize styles at the expense of content, as seen with StyleFormer, or lack sufficient stylistic features, like RAST.

Table 4: User Study Results for Figure 13

	AesPA	RAST	StyTr^2	IEContraAST	StyleFormer	ArtFlow	AdaAttN+RP
Preference ratio ↑	0.01	0.11	0.09	0.06	0.03	0.05	0.65

A.4 Comparison of two representative sparse reconstruction algorithms: FISTA and OMP

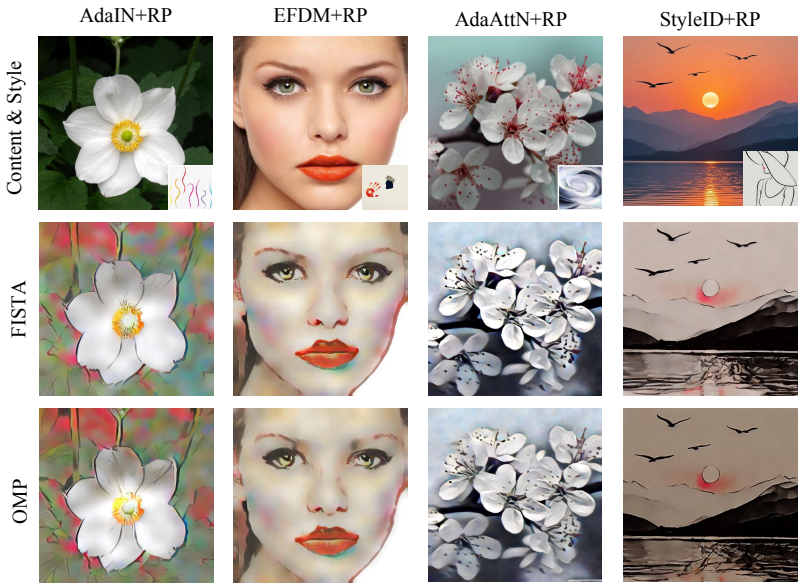

Figure 14: Stylized results derived with two commonly-used sparse reconstruction algorithms: the optimization-based FISTA (Feinman, 2021) vs. the greedy search-based OMP (Lubonja et al., 2024). It is seen that the visual effects of the two algorithms are highly similar.

A.5 Time consumption of our random projection module

In Tables 5, 6, and 7, we present the time consumption for feature encoding, style transfer, and decoding within the conventional networks AdaIN (which shares the same framework as EFDM), AdaAttN, and StyleID. Additionally, we include the time required for the random projection and sparse reconstruction processes within our random projection (RP) module. It is evident that the time overhead introduced by our RP module is very little, particularly when compared to the substantial time savings achieved by eliminating the need for network retraining.

Table 5: Time consumption of AdaIN/EFDM equipped with our RP module

AdaIN/EFDM			Our RP module	
Encoding	Style transfer	Decoding	Random projection	Sparse reconstruction
136ms	18ms	22ms	196ms	203ms

Table 6: Time consumption of AdaAttN equipped with our RP module

AdaAttN			Our RP module	
Encoding	Style transfer	Decoding	Random projection	Sparse reconstruction
872ms	172ms	371ms	2401ms	2303ms

Table 7: Time consumption of StyleID equipped with our RP module

StyleID			Our RP module	
Encoding	Style transfer	Decoding	Random projection	Sparse reconstruction
10533ms	10ms	5603ms	590ms	626ms