SEIT: Structural Enhancement for Unsupervised Image Translation in Frequency Domain

Zhifeng Zhu[1], Yaochen Li[1]*, Yifan Li[1], Jinhuo Yang[1], Peijun Chen[1], Yuehu Liu[2]

[1]School of Software Engineering, Xi'an Jiaotong University
[2]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
z1965761380@stu.xjtu.edu.cn, yaochenli@mail.xjtu.edu.cn,
{3121358033, jinhuo, 3123358029}@stu.xjtu.edu.cn, liuyh@mail.xjtu.edu.cn

Abstract

For the task of unsupervised image translation, transforming the image style while preserving its original structure remains challenging. In this paper, we propose an unsupervised image translation method with structural enhancement in frequency domain named SEIT. Specifically, a frequency dynamic adaptive (FDA) module is designed for image style transformation that can well transfer the image style while maintaining its overall structure by decoupling the image content and style in frequency domain. Moreover, a wavelet-based structure enhancement (WSE) module is proposed to improve the intermediate translation results by matching the high-frequency information, thus enriching the structural details. Furthermore, a multi-scale network architecture is designed to extract the domain-specific information using image-independent encoders for both the source and target domains. The extensive experimental results well demonstrate the effectiveness of the proposed method.

Introduction

Unsupervised image translation is a challenging task that aims to transform images from source domain to target domain without paired training data. In recent years, the development of generative adversarial networks (GANs) (Goodfellow et al. 2020) has led to significant advances in computer vision tasks such as image defogging (Fu et al. 2021), image deraining (Chen et al. 2022), etc. The GAN-based methods have also demonstrated success in accomplishing source-to-target domain translations through adversarial training of generators and discriminators. However, the resulting translated images may exhibit structural distortions and image artifacts.

To address these issues, researchers have made efforts to refine GAN-based techniques for image translation tasks. For example, GCGAN(Fu et al. 2019) ensures the consistency between the input image and the corresponding output by preserving image transformation, such as flipping or rotating. In the work of LPTN(Liang, Zeng, and Zhang 2021), Liang et al. design a framework based on Laplacian decomposition and reconstruction to maintain the image structural information. VSAIT(Theiss et al. 2022) addresses semantic inversion issues by learning inverse mapping in a high-dimensional vector space based on vector symbolic architectures to ensure consistency of source domain content. Although these methods have achieved success in maintaining image structure, the generated results still lack satisfactory style effects.

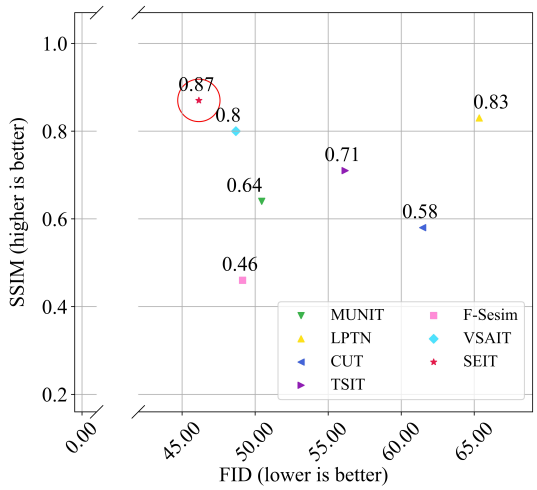

Figure 1: The SSIM vs. FID of different methods on the Day → Night task. The method closest to the top left corner is the best and the red circle represents our proposed SEIT.

To keep both the structure and style well during image translation, we propose an unsupervised image translation method with structural enhancement in the frequency domain named SEIT. The method is based on the GAN framework and consists of a generator and a discriminator. To improve the style of the translation results, we leverage both the source and target domain images as input to the generator, as in previous work (Jiang et al. 2020), to extract more style information during image translation. When image translation is performed in this framework, the content features of the source domain image should be fully preserved and the style features of the target domain image should be fully transferred to the translation result. To meet these requirements, a frequency dynamic adaptive (FDA) module for style conver-

*Corresponding author.

sion is proposed based on the discrete Fourier transform to fully convert the target domain style while maintaining the source domain image content. We also propose a wavelet-based structure enhancement (WSE) module to further improve image detail quality during translation. Based on the above designs, our model is able to maintain the optimal structure and style, as demonstrated in Figure 1, resulting in superior metrics compared with existing methods. The contributions of our work are summarized as follows:

- A novel frequency dynamic adaptive module (FDA) for style conversion is proposed that decouples the image content and style in the frequency domain using the discrete Fourier transform. The content and style information obtained after decoupling are independent of each other, which is conducive to style translation afterwards. During the translation, the designed FDA module can fully capture the style information of the target domain image, while minimizing the loss of source domain image content and transferring the style information of the target domain image.
- A wavelet-based structure enhancement (WSE) module is proposed to enrich the structural details. Through discrete wavelet transformation, the image's style and high-frequency structural information can be separated. Then, the high-frequency features of the intermediate translation results are enhanced using the high-frequency information of the source domain image. The module can enhance the image details without compromising the obtained style information.
- A multi-scale architecture is designed that is independent of both the source and the target domains, which minimizes information loss in encoding and decoding images and extracts domain-specific information during image decoding. The qualitative and quantitative experiments on three different tasks demonstrate that the proposed method achieves state-of-the-art results.

Related Work

Unsupervised Image-to-image Translation

Unsupervised image-to-image translation based on GANs can be divided into one-sided methods and two-sided ones. Zhu et al. (2017) first propose the two-sided method CycleGAN, which introduces the cycle-consistency loss to ensure structural consistency during translation. The methods of DRIT(Lee et al. 2018) and MUNIT(Huang et al. 2018) decompose the latent space into a content-shared space and a domain-specific style space, thus realizing multi-modal image translation. In NICEGAN, Chen et al.(2020) reconsider the role of the discriminator and use the discriminator as part of the generator. Although the cycle-consistency constraint is effective, the underlying two-sided assumption behind the restriction is too absolute. Perfect bidirectional reconstruction is difficult to achieve when the images of one domain have extra information compared to the images of another domain. The one-sided methods aim to ensure that the relationships appearing in the input are reflected in a similar way to the output. Benaim et al. (2017) propose DistanceGAN to constrain the similarity between the distances of the input images and the output images. Park et al. (2020) introduce contrastive learning to constrain the image structure by maximizing the mutual information between the corresponding positions of the original and generated images. The one-sided methods mainly focus on designing various reconstruction losses between the source and generated images. However, the content loss in the flow of the network is ignored. Therefore, these methods are difficult to apply to road scenes to generate images with complete content structures and good details.

Frequency Domain Decomposition

In recent years, many researchers have attempted to combine traditional frequency domain processing methods with deep learning-based image processing. In the field of arbitrary style transfer, Yoo et al. (2019) propose a wavelet pooling strategy to approximate average pooling and its mirror operation upsample pooling, reducing the loss of information when propagating image features in the neural network and improving the quality of images. Zou et al. (2021) design a wavelet transform module to help restore clear high-frequency texture features in image restoration tasks. In the field of unsupervised image translation, Liang et al. (2021) apply Laplacian pyramids to decompose, transform, and reconstruct images to achieve a light network for improving the efficiency of image transformation. Inspired by these methods, we design a wavelet-based structure enhancement module to maintain the texture information and enhance the image structure in the unsupervised image translation.

Approach

Our whole framework adopts a one-sided GAN-based architecture, which consists of a generator and a discriminator. The generator is described in detail next, while the discriminator adopts the same structure as that in (Jiang et al. 2020).

Overall Architecture of the Generator

Figure 2 shows the overall architecture of the generator. Given an image $x \in \mathbb{R}^{3\times h\times w}$ in the source domain and $y \in \mathbb{R}^{3\times h\times w}$ in the target domain, we first feed them to separate encoders to extract multi-scale features. Then the extracted features are fed into the proposed FDA module and WSE module at different layers to fully transfer style information maintaining the overall content and further enhance the structural details. Finally, the image is reconstructed in a multi-scale manner.

In the feature extraction stage, the Conv Block is applied to extract image features, which consists of a 3×3 convolution, instance normalization and relu activation function. And the bilinear interpolation is utilized to perform downsampling. Equipped with the domain-independent encoders, the domain-specific multi-scale features are extracted for further reconstruction. In the image reconstruction stage, the multi-scale source and target domain features are fed into the FDA module and the WSE module to obtain the intermediate features of translation. The features of the deeper layers are upsampled first and then concatenated with the features of the next deeper layers along the channel dimension. Then the image is reconstructed in a coarse-to-fine manner.

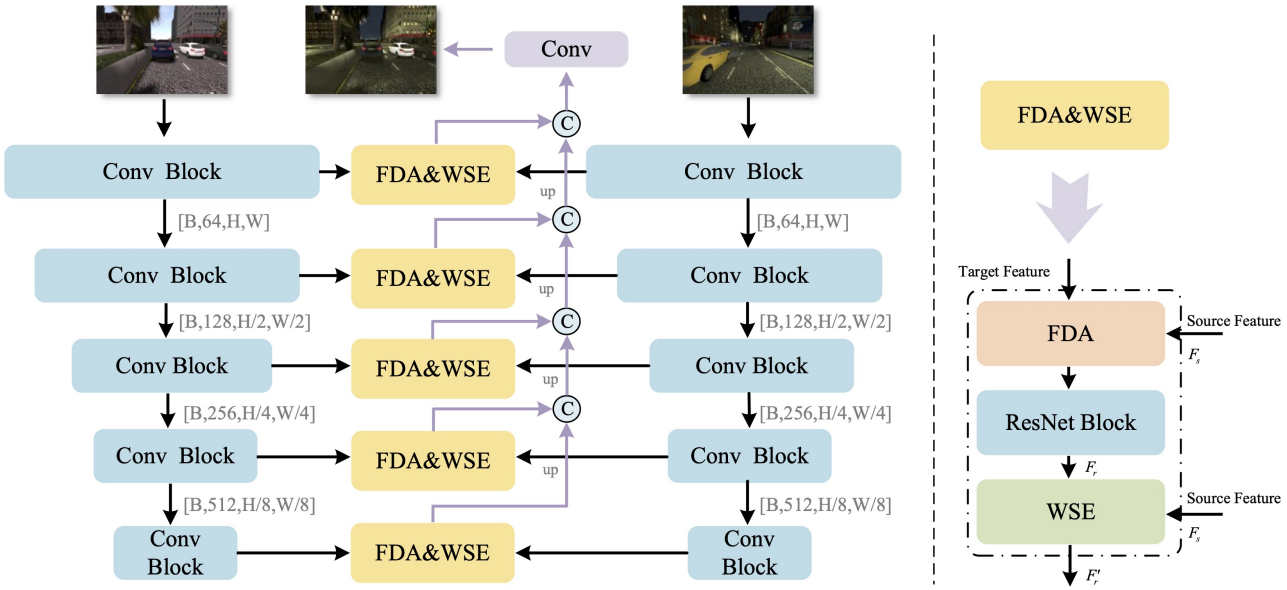

Figure 2: Overall architecture of the generator. The symbols ⓒ and *up* represent channel-wise concatenation and upsample operation.

Frequency Dynamic Adaptive for Style Conversion

We propose a frequency dynamic adaptive (FDA) module for style conversion that can fully capture the style information of the target domain image while minimizing the loss of source domain image content. Figure 3 shows the proposed FDA module, where the amplitude and phase of the source and target domain features can be obtained by discrete Fourier transform, respectively.

For a single channel feature $f(h, w)$ of size $M \times N$ in the spatial domain, the two-dimensional discrete Fourier transform can be expressed as:

$$F(u, v) = \sum_{h=0}^{M-1} \sum_{w=0}^{N-1} f(h, w) e^{-j2\pi(uh/M+vw/N)}, \quad (1)$$

where $F(u, v)$ represents the frequency-domain representation of $f(h, w)$. Then the amplitude and phase of the feature can be obtained by:

$$A = |F(u, v)| = \left[R^2(u, v) + I^2(u, v)\right]^{1/2}, \quad (2)$$

$$P = \Phi(\mathrm{u}, v) = \arctan\left[\frac{I(u, v)}{R(u, v)}\right], \quad (3)$$

where $R(u, v)$ and $I(u, v)$ denote the real and imaginary parts of $F(u, v)$. A denotes the amplitude which contains the style information of the feature, and P represents the phase which contains the content information of the feature.

To obtain the amplitude and phase of a multi-channel feature, each channel is processed independently following the above procedure. After obtaining the amplitude and phase of the feature, we introduce the local style extraction network (LSNet) and the global style extraction network (GSNet), as shown in Figure 3, to extract style information at different scales from the amplitude of the target feature. The LSNet aims to extract the local style modulation parameter W, which first upsamples the input features by 3×3 convolution and then reduces the feature channel dimension by 1×1 convolution to obtain the multi-channel style information. The GSNet aims to extract the global style modulation parameter B, which is distinguished from the LSNet network by the last layer of the adaptive averaging pooling layer. After that, adaptive style conversion is performed as:

$$A_{\hat{F}_s} = W\left(\frac{A_{F_s} - \mu\left(A_{F_s}\right)}{\sigma\left(A_{F_s}\right)}\right) + B, \quad (4)$$

where F_s is the source feature and A_{F_s} denotes its amplitude. $\mu(A_{F_s})$ and $\sigma(A_{F_s})$ are the mean and standard deviation of A_{F_s}. And $A_{\hat{F}_s}$ represents the result of adaptive style conversion. After obtaining the stylized amplitude, the Fourier representation is updated by recombining $A_{\hat{F}_s}$ with the phase of the source feature P_s as:

$$\hat{F}(u, v) = A_{\hat{F}_s} e^{jP_s}. \quad (5)$$

After that, the final style conversion result of each channel $\hat{f}(h, w)$ can be restored using the inverse discrete Fourier transform as:

$$\hat{f}(h, w) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \hat{F}(u, v) e^{j2\pi(uh/M+vw/N)}. \quad (6)$$

And the multi-channel result F_r is formed by combining the single-channel results. Since the FDA module only processes the amplitude of the source feature without changing the phase information, the content of the source feature can be retained to the maximum extent.

Wavelet-based Structure Enhancement

After the adaptive style conversion, the wavelet-based structure enhancement (WSE) module is designed at the feature

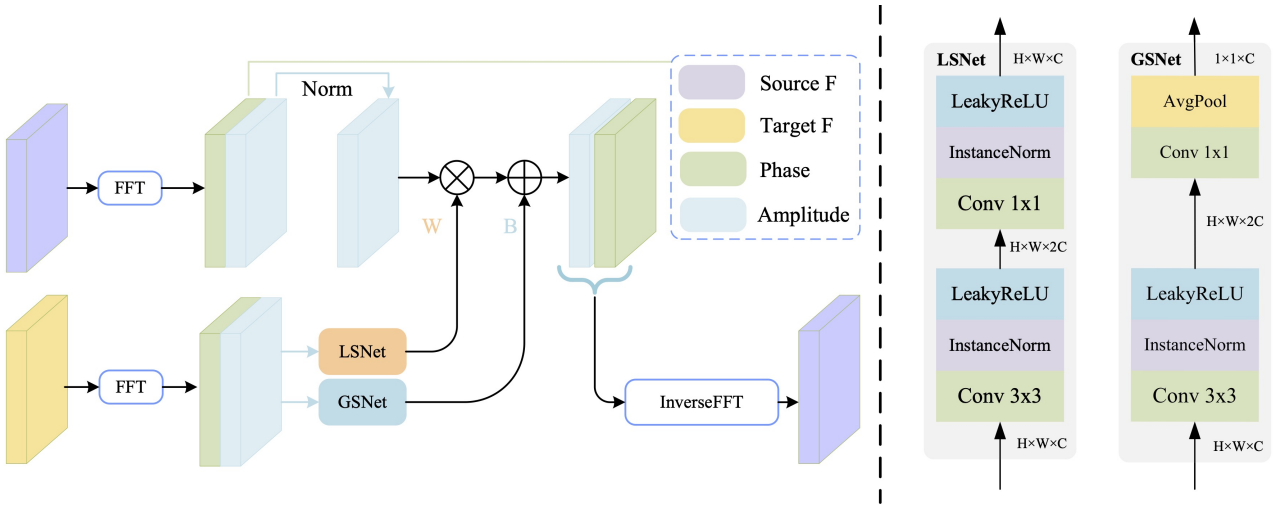

Figure 3: Frequency dynamic adaptive (FDA) style Conversion Module. The symbols $\oplus$ and $\otimes$ represent element-wise addition, element-wise multiplication.

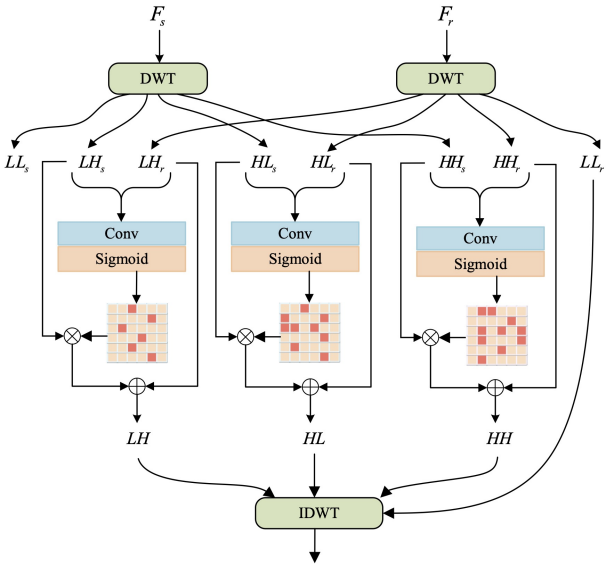

Figure 4: Wavelet-based structure enhancement (WSE). The symbols $\oplus$ and $\otimes$ represent element-wise addition, element-wise multiplication.

level to enrich the structural details. The image features are decomposed into four frequency sub-bands LL, LH, HL and HH after one discrete wavelet transform (DWT). The LL contains the overall image style and retains the global content information. The LH contains high-frequency information in the vertical direction. The HL contains the high-frequency edge information in the horizontal direction. And the HH contains the high-frequency information in the diagonal direction. The Harr wavelet is utilized for the discrete wavelet transform because it has enough ability to portray the image information at different frequencies.

The process of WSE module is shown in Figure 4. The source domain image feature $F_s \in \mathbb{R}^{c\times h\times w}$ and style conversion feature $F_r \in \mathbb{R}^{c\times h\times w}$ are decomposed into four wavelet frequency sub-bands after the DWT. And the principle is defined as:

$$DWT(F_s) = \{{F_s}^{LL}, {F_s}^{LH}, {F_s}^{HL}, {F_s}^{HH}\}, \quad (7)$$

$$DWT(F_r) = \{{F_r}^{LL}, {F_r}^{LH}, {F_r}^{HL}, {F_r}^{HH}\}, \quad (8)$$

where ${F_s}^{LL}$, ${F_s}^{LH}$, ${F_s}^{HL}$ and ${F_s}^{HH}$ represent four frequency sub-bands of F_s. ${F_r}^{LL}$, ${F_r}^{LH}$, ${F_r}^{HL}$ and ${F_r}^{HH}$ represent four frequency sub-bands of F_r. Since the ${F_r}^{LL}$ contains style information while the other three sub-bands contain structural information, we enhance the three high-frequency sub-bands of F_r with the corresponding high-frequency sub-bands of F_s. Let $\odot$ denote the concatenation operation, the enhancement process is defined as:

$$\begin{aligned}
\hat{F}_r^{LH} &= {F_r}^{LH} + {F_s}^{LH} \cdot Sig(Conv(\odot({F_r}^{LH}, {F_s}^{LH}))),\\
\hat{F}_r^{HL} &= {F_r}^{HL} + {F_s}^{HL} \cdot Sig(Conv(\odot({F_r}^{HL}, {F_s}^{HL}))),\\
\hat{F}_r^{HH} &= {F_r}^{HH} + {F_s}^{HH} \cdot Sig(Conv(\odot({F_r}^{HH}, {F_s}^{HH}))).
\end{aligned} \quad (9)$$

After obtaining the enhanced high-frequency sub-bands, the image features are finally reconstructed by inverse discrete wavelet transform (IDWT) by equation (10):

$$\hat{F}_r = IDWT({F_r}^{LL}, \hat{F}_r^{LH}, \hat{F}_r^{HL}, \hat{F}_r^{HH}). \quad (10)$$

Note that the low-frequency sub-band ${F_r}^{LL}$, which contains the stylized information of F_r, is not changed in the enhancement process. Therefore, the module can enhance the image details without compromising the obtained style information.

Loss Function

The loss functions of our method contain the adversarial loss L_{adv}, the perceptual loss L_p and the feature matching loss L_{fm}.

Adversarial Loss. The adversarial loss is applied to guide the generator to translate the style to the target domain, which is represented as:

$$L_{adv} = \mathbb{E}_{x\sim P_{data}(X)}(D(x)) + \mathbb{E}_{x\sim P_{data}(X),y\sim P_{data}(Y)}(1 - D((G(x,y))). \quad (11)$$

Perceptual loss. The perceptual loss(Johnson, Alahi, and Fei-Fei 2016) is used to make the content of the translation result similar to the input source image and is defined as:

$$L_p = E_{x\sim P_{data(X)},y\sim P_{data(Y)}}[\alpha_j \|\Phi_j(G(x,y)) - \Phi_j(x)\|_1], \quad (12)$$

where Φ denotes the pre-trained VGG(Simonyan and Zisserman 2014) network, $\Phi_j(x)$ denotes the feature after the activation function of the j_{th} layer and α_j is the weight of the j_{th} layer. We use five layers of intermediate features from the pre-trained VGG network, namely $relu1\_1$, $relu2\_1$, $relu3\_1$, $relu4\_1$, $relu5\_1$, whose corresponding weights are 1/32, 1/16, 1/8, 1/4, 1, respectively.

Feature Matching Loss. The feature matching loss is used to match intermediate layer features of multi-scale discriminator:

$$L_{fm} = E_{x\sim P_{data(X)},y\sim P_{data(Y)}}[\left\|D^i(G(x,y)) - D^i(y)\right\|_1], \quad (13)$$

where $D^i(\cdot)$ is the result of the discriminator's i_{th} layer.

Total Loss. The total loss is a combination of the aforementioned ones:

$$L_{total} = \lambda_{adv}L_{adv} + \lambda_p L_p + \lambda_{fm}L_{fm}. \quad (14)$$

where λ_{adv} is the weight of adversarial loss, λ_p is the weight of the perceptual loss and λ_{fm} is the weight of the feature matching loss.

Experiments

We conduct experiments on the cross-time translation, cross-weather translation and cross-dataset translation tasks, respectively. The detailed experimental setup is as follows.

Experimental Setup

Datasets The datasets we use include SYNTHIA(Ros et al. 2016), GTA5(Richter et al. 2016), Cityscapes(Cordts et al. 2015) and BDD(Yu et al. 2020). we conduct the Day→Night translation on SYNTHIA, the Sunny→Cloudy translation on BDD, and the cross-domain translation on GTA5 and Cityscapes.

Evaluation Metrics SSIM(Wang et al. 2004) is applied to measure the structural similarity of the translation result to the source domain image, which combines three types of information: brightness, structure, and contrast of the image. The value of SSIM is closer to 1 when the structure of both is more similar.

FSIM(Zhang et al. 2011) is also applied to measure the structural similarity of the translation result to the source domain image. It combines the Phase Congruency and Gradient Magnitude of the image. The value of FSIM is closer to 1 when the structure of both is more similar.

FID(Heusel et al. 2017) is utilized for measuring the similarity of the distribution between the target domain and the translation results. The lower the FID, the closer of the two distributions are.

Training Details All experiments are conducted on a single RTX 3090 GPU. The batch size is set to 1. We use the Adam optimizer with β_1 = 0.5 and β_2 = 0.999. The initial learning rate is set to 0.0002 and the step decay learning strategy is used, with the learning rate decaying to half of the original learning rate every 5 epochs. The model is trained for 100 epochs. Following previous work, the loss weight in equation 14 is set to 1.0, 2.0, and 1.0, respectively.

Baselines We compare with GAN-based image translation methods including MUNIT (Huang et al. 2018), LPTN (Liang, Zeng, and Zhang 2021), CUT (Park et al. 2020), TSIT (Jiang et al. 2020), F-Sesim (Zheng, Cham, and Cai 2021) and VSAIT (Theiss et al. 2022).

Qualitative Comparisons

The visual results of our method on three different tasks are shown in Figures 5-7. For each task, we compare two visual examples with zoomed-in local details to evaluate the performance of each method in preserving image details and transferring style. As shown in Figures 5-7, the MUNIT (Huang et al. 2018), CUT (Park et al. 2020), TSIT (Jiang et al. 2020) and F-Sesim (Zheng, Cham, and Cai 2021) suffer from structural distortions and artifacts. Although LPTN (Liang, Zeng, and Zhang 2021) and VSAIT (Theiss et al. 2022) can retain the structural features better (see Figures 5 and 6), they have limited style transfer ability. VSAIT (Theiss et al. 2022) produces a satisfactory result when the source and target domains share similar styles (see Figure 7), whereas it still lacks style transfer ability. In contrast, our method can faithfully preserve the structural details of the source image and fully capture the style of the target domain. Our method outperforms other methods for high-quality image translation in both the overall quality and the zoomed-in details.

Quantitative Comparisons

The quantitative results of our method on three tasks are shown in Table 1. Our method achieves the best performance among all compared methods on all tasks, demonstrating its ability to generate images with good structure and style. Notably, our method has a significant improvement on SSIM (Wang et al. 2004) and FSIM (Zhang et al. 2011) while maintaining a clear advantage on FID (Heusel et al. 2017) for the GTA5 → Cityscapes task. The above comparison indicates the efficiency of our method in preserving the content structure and transferring the style of the input images.

Ablation Study

Qualitative comparisons To verify the effectiveness of our method, ablation experiments are conducted on both the single-scale and the multi-scale architecture. The qualitative results are shown in Figure 8. The result of the single-scale architecture in the yellow box (see Figure 8(a)) fails to restore the wires next to the streetlights in the original image

Method	SYNTHIA Day → Night			BDD Sunny → Cloudy			GTA5 → CityScapes		
	SSIM ↑	FSIM ↑	FID ↓	SSIM ↑	FSIM ↑	FID ↓	SSIM ↑	FSIM ↑	FID ↓
MUNIT	0.64	0.81	50.45	0.88	0.93	41.69	0.63	0.79	105.15
LPTN	0.83	0.88	65.34	0.92	0.94	46.37	0.77	0.89	99.65
CUT	0.58	0.77	61.74	0.81	0.84	58.33	0.55	0.70	98.08
TSIT	0.71	0.80	56.16	0.89	0.90	44.24	0.70	0.82	98.61
F-Sesim	0.46	0.84	49.14	0.79	0.91	42.55	0.55	0.74	83.21
VSAIT	0.80	0.90	48.68	0.89	**0.96**	42.50	0.58	0.86	82.77
Ours	**0.87**	**0.91**	**46.15**	**0.94**	**0.96**	**40.74**	**0.93**	**0.95**	**82.55**

Table 1: Quantitative comparisons on three tasks. Bolded represents the best and underline indicates the second best.

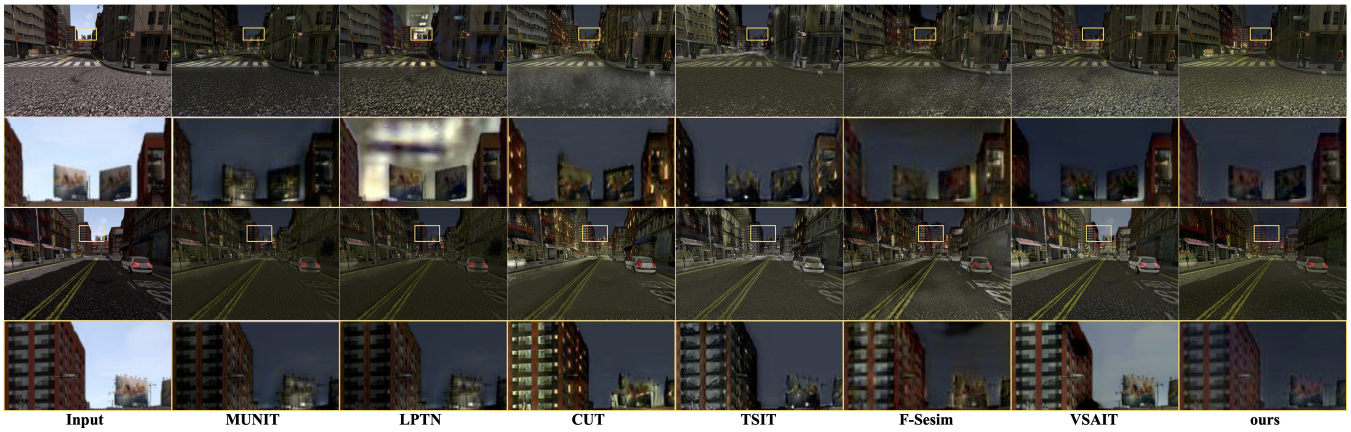

Figure 5: Comparison of existing methods on the SYNTHIA Day → Night task.

S-S	M-S	FDA	WSE	SSIM ↑	FSIM ↑	FID ↓
✓				0.54	0.80	54.70
✓		✓		0.67	0.85	49.50
✓			✓	0.64	0.84	48.26
✓		✓	✓	0.70	0.86	47.46
	✓			0.78	0.85	46.68
	✓		✓	0.80	0.87	46.44
	✓	✓		0.83	0.88	46.99
	✓	✓	✓	**0.87**	**0.91**	**46.15**

Table 2: Quantitative comparison results of ablation studies. S-S represents single-scale and M-S represents Multi-scale.

and the style is chaotic. In the red box, the translation result fails to restore the pedestrians and loses the structure information. These problems are caused by severe information loss in encoding and decoding. In Figure 8(b)(c), it can be seen that the visual effects of these two parts are significantly improved. Compared to Figure 8(b)(c), the results in Figure 8(d) are relatively better in stylization effect and structure maintenance, which demonstrates that our method can improve performance on the single-scale architecture.

In the multi-scale architecture results, as shown in Figure 8(e), the structure of the translation results is more complete and clearer than that of the single-scale results, which is attributed to the multi-scale feature extraction capability. However, the stylization effect is still unsatisfactory. In Figure 8(f), by adding the FDA module, both the global content and the stylization effect are improved compared to the multi-scale method. And in Figure 8(g), the detailed structure is enhanced by adding the WSE module. When all of our proposed modules are applied, the result shown in Figure 8(h) achieves the best in terms of both style effect and structure maintenance, which proves the effectiveness of our method.

Quantitative Comparisons The quantitative results of the ablation experiments are shown in Table 2. The results of the single-scale architecture are the worst, and the SSIM (Wang et al. 2004) and FSIM (Zhang et al. 2011) are improved by adding FDA or WSE to the network. The best results are obtained by adding both the proposed modules. The baseline results of the multi-scale architecture outperform the optimal results under the single-scale architecture in terms of SSIM (Wang et al. 2004) and FID (Heusel et al. 2017), which demonstrate the effectiveness of the proposed multi-scale architecture. The quantitative results are further improved after adding FDA or WSE. The best results are achieved when all modules are applied, which is consistent with the qualitative experimental results.

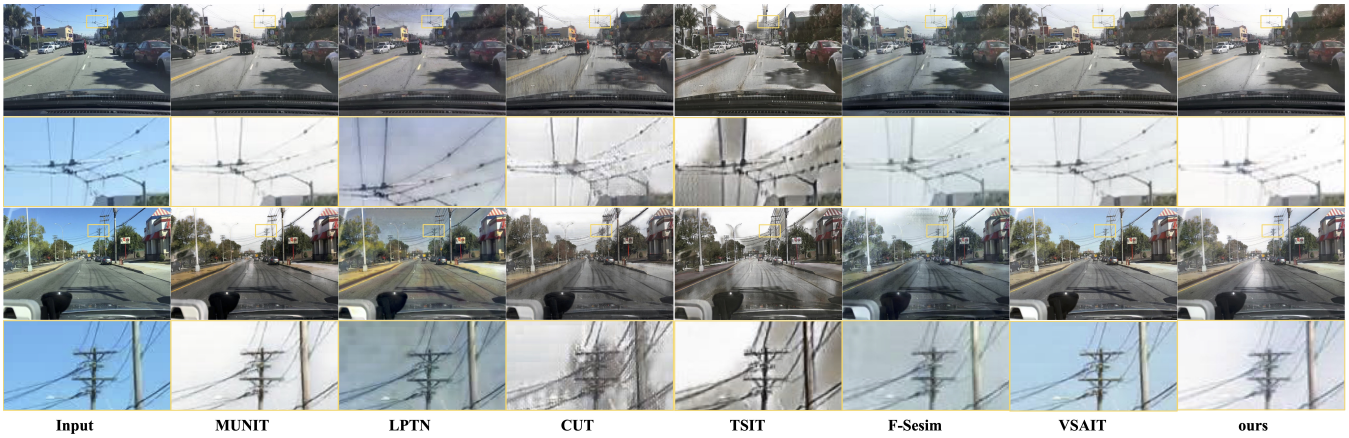

Figure 6: Comparison of existing methods on the BDD Sunny → Cloudy task.

Figure 7: Comparison of existing methods on the GTA5 → CityScapes task.

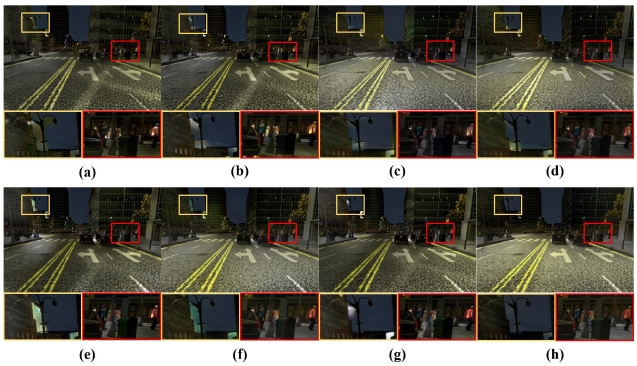

Figure 8: Qualitative comparison results of ablation studies. (a) Single-scale. (b) Single-scale+FDA. (c) Single-scale+WSE. (d) Single-scale+FDA+WSE. (e) Multi-scale. (f) Multi-scale+FDA. (g) Multi-scale+WSE. (h) Multi-scale+FDA+WSE.

Conclusion and Future Work

In this work, we introduce an unsupervised image translation method with structural enhancement in frequency domain named SEIT. It is built on the GAN framework with the proposed FDA and WSE modules. We take advantage of the image features in the frequency domain to preserve the source content feature during style conversion and further enhance the structural details. The multi-scale architecture is applied to minimize information loss during translation and domain-specific features are extracted using image-independent encoders for the source and target domains. Our method outperforms existing methods qualitatively and quantitatively. Thanks to the effective decoupling of content and style, our method can be extended to the multi-modal image translation tasks to explore its performance on more domains.

Acknowledgements

This work was supported by the National Key Research and Development Project of New Generation Artificial Intelligence of China under Grant 2018AAA0102504, and Key R&D Plan of Shaanxi Province under grant number 2022GY-080.

References

Benaim, S.; and Wolf, L. 2017. One-sided unsupervised domain mapping. *Advances in Neural Information Processing Systems*, 30.

Chen, R.; Huang, W.; Huang, B.; Sun, F.; and Fang, B. 2020. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8168–8177.

Chen, X.; Pan, J.; Jiang, K.; Li, Y.; Huang, Y.; Kong, C.; Dai, L.; and Fan, Z. 2022. Unpaired deep image deraining using dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017–2026.

Cordts, M.; Omran, M.; Ramos, S.; Scharwächter, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2015. The cityscapes dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop on the Future of Datasets in Vision*, volume 2.

Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Zhang, K.; and Tao, D. 2019. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2427–2436.

Fu, M.; Liu, H.; Yu, Y.; Chen, J.; and Wang, K. 2021. DW-GAN: A discrete wavelet transform GAN for Nonhomogeneous Dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 203–212.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, 172–189.

Jiang, L.; Zhang, C.; Huang, M.; Liu, C.; Shi, J.; and Loy, C. C. 2020. TSIT: A simple and versatile framework for image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, 206–222. Springer.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, 694–711. Springer.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision*, 35–51.

Liang, J.; Zeng, H.; and Zhang, L. 2021. High-resolution photorealistic image translation in real-time: A laplacian pyramid translation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9392–9400.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive learning for unpaired image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, 319–345. Springer.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision*, 102–118. Springer.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3234–3243.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Theiss, J.; Leverett, J.; Kim, D.; and Prakash, A. 2022. Unpaired Image Translation via Vector Symbolic Architectures. In *Proceedings of the European Conference on Computer Vision*, 17–32. Springer.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Yoo, J.; Uh, Y.; Chun, S.; Kang, B.; and Ha, J.-W. 2019. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9036–9045.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2636–2645.

Zhang, L.; Zhang, L.; Mou, X.; and Zhang, D. 2011. FSIM: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8): 2378–2386.

Zheng, C.; Cham, T.-J.; and Cai, J. 2021. The spatially-correlative loss for various image translation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16407–16417.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2223–2232.

Zou, W.; Jiang, M.; Zhang, Y.; Chen, L.; Lu, Z.; and Wu, Y. 2021. SDWNet: A straight dilated network with wavelet transformation for image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1895–1904.