Multidisciplinary Journal of Engineering and Technology
ISSN: 3007-6900
DOI: https://doi.org/10.61784/mjet3010

ENHANCED UNSUPERVISED IMAGE-TO-IMAGE TRANSLATION USING CONTRASTIVE LEARNING AND HISTOGRAM OF ORIENTED GRADIENTS

WanChen Zhao, WenHan Wang, Cheng Zhang, XiaoLei Qu*
College of Instrument Science and Optoelectronic Engineering, Beihang University, Beijing 100191, China.
Corresponding Author: XiaoLei Qu, Email: quxiaolei@buaa.edu.cn

Abstract: Image-to-Image Translation is a vital area of computer vision that focuses on transforming images from one visual domain to another while preserving their core content and structure. However, this field faces two major challenges: first, the data from the two domains are often unpaired, making it difficult to train generative adversarial networks effectively; second, existing methods tend to produce artifacts or hallucinations during image generation, leading to a decline in image quality. To address these issues, this paper proposes an enhanced unsupervised image-to-image translation method based on the Contrastive Unpaired Translation (CUT) model, incorporating Histogram of Oriented Gradients (HOG) features. This novel approach ensures the preservation of the semantic structure of images, even without semantic labels, by minimizing the loss between the HOG features of input and generated images. The method was tested on translating synthetic game environments from GTA5 dataset to realistic urban scenes in cityscapes dataset, demonstrating significant improvements in reducing hallucinations and enhancing image quality.
Keywords: Image-to-image translation; Photorealism; GANs

1 INTRODUCTION

Image-to-Image Translation has emerged as a pivotal area of research within computer vision and machine learning, focusing on the intricate task of transforming images from one visual domain to another while preserving their inherent content and structure. This capability has wide-ranging applications, including but not limited to converting sketches into lifelike photographs [1], colorizing grayscale images [2], and altering scenes from day to night [3]. The overarching goal of image-to-image translation is to learn a robust mapping between two distinct visual domains, enabling seamless transitions that retain the essence of the original imagery.

Generative Adversarial Networks (GANs) have become the cornerstone of this transformation process, offering a powerful deep learning framework tailored for image generation tasks. Early breakthroughs in this field predominantly relied on supervised learning with paired datasets, where each image in the source domain has a corresponding counterpart in the target domain. A seminal work in this domain is pix2pix [4], which employs conditional GANs to learn a direct mapping from input images to output images by leveraging paired training data. Despite its effectiveness, this approach is constrained by the necessity of paired datasets, which are often difficult, time-consuming, and costly to obtain, particularly in complex or large-scale applications.

To circumvent the challenges associated with paired datasets, researchers have developed innovative methods that facilitate translation between unpaired datasets. CycleGAN [5] stands out as a landmark approach, utilizing dual generators (G and F) and dual discriminators. Generator G transforms images from style X to style Y, while generator F performs the reverse operation. The discriminators are tasked with differentiating between real and generated images in both styles. Given the absence of paired data, adversarial loss alone cannot guarantee accurate translation, prompting the introduction of a cycle-consistency loss. This loss ensures that an image can be reconstructed after passing through both generators (i.e., $F(G(X)) \approx X$ or $G(F(Y)) \approx Y$), thereby maintaining the content and structural integrity of the images. While CycleGAN has proven to be a versatile and effective solution, it is not without its limitations, including the high computational burden of training two sets of generators and discriminators and the tendency to produce artifacts in the generated images.

The inherent weakness of cycle-consistency loss, which can lead to image artifacts or hallucinations, has spurred the development of more refined techniques. StyleGAN [1] , for instance, introduces a mapping network that projects the input latent vector into a style vector space, allowing for fine-grained control over the generative features and significantly reducing instability during training. This approach not only enhances the aesthetic quality of the generated images but also allows for greater manipulation of the image's stylistic elements. Sem-GAN [6] further advances this concept by integrating semantic labels into the training process, ensuring that the generated images are both visually convincing and semantically consistent, thus addressing the challenge of maintaining meaningful content across translations.

Attention mechanisms have also played a crucial role in advancing image-to-image translation by addressing the challenge of capturing high-level semantic information. Attention GAN [7], for example, uses attention-guided generators to create attention masks, which are then combined with the input image to generate high-quality target images. This selective focus on critical image regions helps to preserve important details and improve the overall

translation quality. Conversely, SPA-GAN [8] enhances the discriminator's ability to discern fine details by incorporating attention mechanisms, which in turn improves the generator's capacity to produce more realistic images.
In addition to these methods, Unsupervised Image-to-Image Translation (UNIT) [9] leverages a shared latent space assumption, mapping images from different domains into a common feature space. This approach provides an effective solution for more complex unsupervised cross-domain image translation tasks, enabling the model to learn shared features that are domain-invariant. Similarly, CUT [10] employs contrastive learning to maximize the similarity between source and target domain images in the latent space while preserving their distinct characteristics. This technique has proven particularly effective in maintaining the unique aspects of each domain while ensuring a coherent translation.

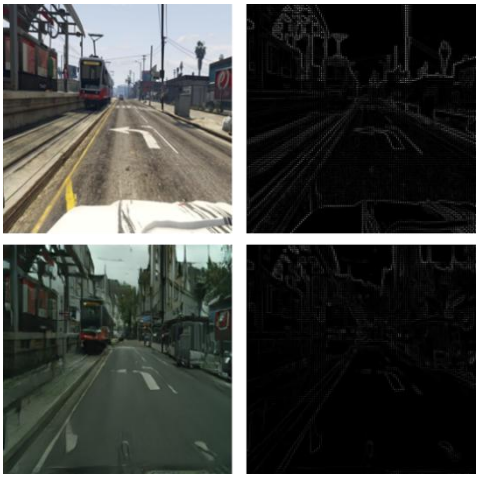

Figure 1 Generation of Image Hallucinations and Their HOG Feature Maps. The First Row Shows the Original Images along with their HOG Feature maps, while the Second Row Displays the Images Generated by the CUT Model and their Corresponding HOG Feature Maps

Despite the significant advancements achieved by these methods, several challenges persist. For instance, while the translation quality of the main subject in an image is often high, the quality of background translation can be inconsistent, leading to unwanted artifacts or hallucinations. This issue is particularly problematic in photorealism tasks, such as translating game screenshots into real-world images, where even minor semantic errors—such as misinterpreting the sky as a building—can result in glaring inaccuracies. Moreover, the reliance on semantic supervision poses a practical limitation, as many real-world datasets lack the necessary semantic labels, rendering these approaches less effective in unsupervised settings.
To address these challenges, we have developed a novel unsupervised image-to-image translation method based on the CUT model, incorporating a unique loss function grounded in Histogram of Oriented Gradients (HOG) features [11]. HOG features divide an image into small cells and compute histograms of gradient directions within each cell, providing robustness to variations in lighting and geometric transformations. By calculating the HOG features of both the input and generated images and computing the L2 loss between them, our method ensures that the semantic structure of the generated image is preserved, even in the absence of semantic label supervision. This approach not only mitigates the issue of hallucinations but also enhances the overall quality and fidelity of the translated images.
To validate the effectiveness of our proposed method, we conducted a series of experiments on image translation from GTA5 [12] to Cityscapes [13], a challenging task that involves translating synthetic game environments into realistic urban scenes. The experimental results, supported by comprehensive visualizations, demonstrate that our method effectively reduces hallucinations in image generation, significantly improving the quality and realism of the translated images. Our approach represents a meaningful step forward in unsupervised image-to-image translation, offering a robust solution to some of the most persistent challenges in this field.

2 RELATED WORK

2.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) [14], introduced by Goodfellow et al. in 2014, have revolutionized the field of generative modeling by presenting a novel framework where two neural networks, a generator and a discriminator, engage in a competitive training process. The generator aims to create images that are indistinguishable from real data, while the discriminator's goal is to differentiate between real and generated images. This adversarial interplay leads to the generator producing increasingly realistic images, and the discriminator becoming more adept at detecting fakes. The success of this dynamic has led to the development of numerous GAN variants, each designed to address specific challenges or enhance the model’s capabilities.
Among the earliest significant improvements was the Deep Convolutional GAN (DCGAN) [15], which introduced a more stable architecture for GANs by leveraging convolutional layers and replacing deterministic pooling functions

with strided convolutions. This architectural shift not only stabilized training but also improved the visual quality of the generated images, making DCGANs a foundational model for many subsequent GAN-based studies.

Conditional GANs (cGANs) [16], another pivotal advancement, extended the GAN framework by incorporating auxiliary information such as class labels, enabling the generation of images conditioned on specific input data. This capability has made cGANs highly effective for tasks requiring controlled image synthesis, such as image-to-image translation, where the generated image must adhere to specific characteristics dictated by the input.

StyleGAN [1], developed by Karras et al., represents a significant leap forward in GAN research. By introducing a new generator architecture that allows for the control of style at various levels of the synthesis process, StyleGAN enables the fine-tuning of generated images in ways that were previously impossible. This model has set new standards for photorealism in image synthesis, and its innovations, such as the mapping network and style mixing regularization, have been widely adopted in subsequent works.

Research into GANs has also delved deeply into the theoretical underpinnings and optimization strategies to overcome challenges like mode collapse and unstable training dynamics. Techniques such as Wasserstein GANs (WGANs) [17], which use the Earth Mover's Distance (Wasserstein distance) as a loss function, have provided more robust convergence properties and reduced the risk of mode collapse. Additionally, the incorporation of spectral normalization has been shown to stabilize the training of GANs by constraining the Lipschitz constant of the discriminator, further contributing to the reliability of GAN training processes [18].

Furthermore, GANs have found extensive application in the domain of image-to-image translation, significantly advancing the capabilities of this field. For instance, GAN-based models have been employed in translating low-resolution images to high-resolution counterparts [19], enhancing the quality and details in tasks like super-resolution. Additionally, GANs have been pivotal in domain adaptation, where models are trained to translate images from one domain (e.g., sketches) to another (e.g., photorealistic images) while preserving essential content. Their flexibility has also been demonstrated in cross-modal translation tasks, such as converting grayscale images to color [1] or transforming aerial images into map-like representations [20]. The versatility of GANs in handling diverse image translation tasks underlines their profound impact on both academic research and practical applications, making them a cornerstone technology in the advancement of computer vision.

2.2 Image-to-Image Translation

Image-to-Image Translation is a critical research area in computer vision, focused on transforming an image from one domain into another while preserving its core content and structure. The field has been significantly advanced by the application of Generative Adversarial Networks (GANs), with the pix2pix framework by Isola et al. serving as a landmark model. Pix2pix demonstrated the power of conditional GANs (cGANs) in supervised image translation tasks, where paired datasets—each containing an input and a corresponding target image—are used to train the model. This framework has been particularly successful in tasks such as image inpainting, where missing parts of an image are filled in, and in style transfer, where an image's artistic style is altered while maintaining its underlying structure.

However, the reliance on paired datasets poses significant limitations, as collecting such data can be challenging in many scenarios. To address this, Zhu et al. proposed CycleGAN, which introduced cycle consistency as a key concept to enable translation between two domains without requiring paired data. CycleGAN's ability to learn mappings between domains using unpaired datasets has made it widely applicable in various tasks, including artistic style transfer and domain adaptation, where direct correspondences between images are unavailable.

Following CycleGAN, numerous methods have been developed to enhance the performance and applicability of image-to-image translation models. For example, DiscoGAN [21] and DualGAN [22] introduced additional constraints, such as identity loss and dual learning mechanisms, to improve the stability and quality of translations in unpaired settings. These models have contributed to a better understanding of how to maintain content consistency while achieving the desired stylistic transformation.

Multi-Modal UNsupervised Image-to-Image Translation (MUNIT) [23] and StarGAN [24] represent significant strides in handling multi-domain and multi-modal translation. MUNIT allows for the separation of content and style, enabling the generation of diverse outputs from a single input image by recombining different content and style codes. This flexibility has been particularly useful in applications where multiple plausible outputs exist for a given input, such as in image editing and creative content generation. StarGAN, on the other hand, unifies the process of image translation across multiple domains within a single model, providing a scalable solution for tasks requiring multi-domain transformations.

Recent innovations like the Contrastive Unpaired Translation (CUT) model [10] have introduced contrastive learning into the image-to-image translation pipeline. CUT maximizes mutual information between the input and output domains, enabling high-quality translation even when only a single image from each domain is available. This approach has proven particularly effective in one-sided translation tasks, where traditional methods might struggle due to limited data availability.

Moreover, Stephan R. Richter et al.'s work on utilizing intermediate representations for adversarial training has highlighted the importance of perceptual supervision at multiple levels within a neural network [25]. By enforcing adversarial constraints at different perceptual layers, this method has achieved state-of-the-art results in tasks requiring fine-grained control over the translation process, such as semantic segmentation and high-resolution image synthesis.

In addition to these advancements, researchers have also explored ways to improve the interpretability and control of image-to-image translation models. Approaches like disentanglement learning, where models learn to separate different factors of variation within the data, have allowed for more precise control over the generated outputs [26]. This has opened up new possibilities for user-guided image translation, where specific aspects of the image can be altered according to user-defined parameters, thereby enhancing the practicality and usability of these models in real-world applications.

These developments illustrate the ongoing progress in image-to-image translation research, driven by the continuous refinement of GAN-based methods and the exploration of new learning paradigms that address both practical challenges and theoretical concerns in the field.

3 METHOD

Our overall architecture is based on the structure of CUT. Building upon the CUT model, we further introduce HOG feature loss to supervise the training of the generator. By leveraging the robustness of HOG features against style variations, we ensure that the generated realistic images remain consistent with the original simulated images in terms of content and spatial structure, thereby achieving a better realistic effect for the simulated images. A schematic diagram of this structure is shown in Figure 2.

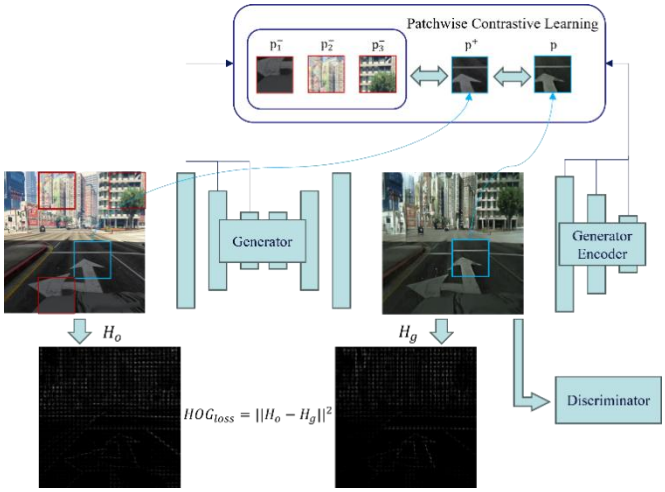

Figure 2 The Architecture of our Proposed Method

3.1 CUT Model

CUT is specifically designed to address the challenges of unpaired image translation, especially when paired samples are unavailable. Traditional unpaired image translation methods often rely on cycle consistency loss, which necessitates translating images back and forth between the source and target domains. However, this approach has limitations in preserving image details and content structure and incurs significant computational costs. CUT overcomes these challenges by introducing contrastive learning and a one-sided translation strategy.

The core idea of CUT is to preserve content consistency by maximizing the mutual information between corresponding local regions (patches) of the input and output images. Specifically, CUT employs a contrastive loss mechanism, which pulls positive samples (corresponding patches) closer together while pushing negative samples (non-corresponding patches) farther apart. This ensures that the generated images retain the local features of the original images. To achieve this, CUT introduces a patch-based discriminator that evaluates local regions of the image rather than the image as a whole. This allows the model to learn domain translation at a global scale while maintaining fine-grained details at a local scale.

In its one-sided translation strategy, CUT focuses solely on translating from the source domain to the target domain, eliminating the need for reverse mapping. This strategy significantly reduces computational overhead and avoids potential artifacts introduced by cycle consistency constraints. Additionally, CUT employs a technique known as "contrastive self-supervision," which enhances model stability and translation quality by introducing self-supervised mechanisms in the absence of paired data. This allows CUT to generate high-quality translations even with unidirectional datasets.

The CUT method is particularly suitable for scenarios where paired data is difficult to obtain, such as artistic style transfer and medical image processing. In artistic style transfer, CUT can preserve the content of the original image while achieving nuanced style transformations. In medical imaging, it can effectively enhance low-quality images to high-quality ones, playing a critical role in medical diagnosis.

Overall, the CUT method, by combining contrastive learning, one-sided translation, and self-supervision mechanisms, significantly enhances the effectiveness of unpaired image translation. Its robust ability to retain local details and efficient computational performance make it a highly promising tool in the field of image-to-image translation.

3.2 HOG Loss

As depicted in Figure 1, while the CUT model utilizes a patch-based discriminator to ensure that the images generated by the generator retain fine details and achieve overall stability, the problem of hallucination—where the generated images may contain unrealistic or extraneous features—still persists. To address this issue and further tighten the constraints on the generated images, we have incorporated HOG (Histogram of Oriented Gradients) loss into the training process of the generator.

The introduction of HOG loss is strategically aimed at harnessing the robustness of HOG features, which are particularly effective in capturing and preserving structural details and edge orientations despite variations in style. By aligning the HOG features of the generated images with those of the input images, this loss function promotes structural invariance between the source and target domains. This alignment helps in mitigating hallucinations, ensuring that the generated images are not only stylistically transformed but also maintain the structural integrity of the original content.

HOG loss operates by comparing the HOG features extracted from both the generated and reference images. It encourages consistency in these features, penalizing deviations that could result in visual artifacts or inconsistencies. This method enhances the generator's capability to produce more accurate and realistic outputs, thereby reducing the likelihood of hallucinated elements.

The calculation of HOG loss is as follows:

$$\mathrm{HOG}_{\mathrm{loss}} = \left|\left|\mathrm{H}_{\mathrm{o}} - \mathrm{H}_{\mathrm{g}}\right|\right|^{2} \# \tag{1}$$

where H_{o} denotes the HOG features of the original images and H_{g} denotes the HOG features of the generated images.

4 EXPERIMENT AND RESULTS

4.1 Dataset

We conduct our experiments using the following two datasets:

The GTA5 dataset [12] is a synthetic dataset tailored for computer vision tasks. It is derived from the popular video game Grand Theft Auto V (GTA V), where the in-game urban environment is utilized to simulate real-world street scenes. This dataset offers high-resolution images with pixel-level annotations that include various urban objects such as buildings, vehicles, roads, and sidewalks. The Cityscapes dataset [13] is a real-world dataset specifically curated for visual understanding tasks in urban settings. It is extensively used in autonomous driving research. This dataset comprises high-resolution images captured from 50 different cities across Germany, with detailed annotations covering 30 classes, 19 of which are commonly employed for evaluation. Both datasets consist of 2,500 images each.

Figure 3 Illustration of the Datasets. The First Row Represents the GTA5 Dataset, while the Second Row Depicts the Cityscapes Dataset

4.2 Evaluation Metrics

In this paper, we use three evaluation metrics—the Inception Score (IS) [27], Kernel Inception Distance (KID) [28], and Fréchet Inception Distance (FID) [29]—to assess the realism of generated images. Each of these metrics provides unique insights into different aspects of image generation performance.

The Inception Score (IS) evaluates both the clarity and diversity of generated images. It measures how well the generated images can be classified into distinct categories and how diverse these generated samples are. Specifically, IS uses a pre-trained Inception v3 model [30] to compute the Kullback-Leibler (KL) divergence between the conditional

label distribution p(y | x) for each generated image x and the marginal label distribution p(y) across multiple generated images. A higher IS indicates that the images are not only of high quality but also exhibit a diverse range of classes.

$$\mathrm{IS} = \exp\left(\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{KL}(\mathrm{p}(\mathrm{y}|\mathrm{x_i})||\mathrm{p}(\mathrm{y}))\right)\# \tag{2}$$

Where KL denotes the Kullback-Leibler divergence and N denotes the number of samples.

The Kernel Inception Distance (KID) measures the similarity between the feature distributions of real and generated images. Unlike IS, which focuses on classification, KID uses a kernel function to compare the distributions of features extracted from a pre-trained Inception model. The metric calculates the Maximum Mean Discrepancy (MMD) between these feature distributions, providing an unbiased estimate of the distance between the real and generated image distributions. A lower KID score suggests that the generated images closely match the real image distribution in feature space.

$$\mathrm{KID} = \frac{1}{\mathrm{M}^2}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{M}} \varphi(\mathrm{x_i})\varphi(\mathrm{x_j}) - \frac{2}{\mathrm{MN}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{M}} \varphi(\mathrm{x_i})\varphi(\mathrm{y_j}) + \frac{1}{\mathrm{N}^2}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{M}} \varphi(\mathrm{y_i})\varphi(\mathrm{y_j})\,\# \tag{3}$$

Where φ represents the feature map, and M and N denote the number of real and generated samples, respectively.

The Fréchet Inception Distance (FID) assesses the distance between the feature distributions of real and generated images. It calculates the Fréchet distance (or Wasserstein distance) between the Gaussian distributions fitted to the feature vectors of both real and generated images. The metric takes into account both the mean and covariance of these distributions. A lower FID score indicates that the generated images have feature distributions that are more similar to those of real images, reflecting higher image quality and better alignment with the real image distribution.

$$\mathrm{FID} = \left|\left|\mu_\mathrm{r} - \mu_\mathrm{g}\right|\right|_2^2 + \mathrm{Tr}\left(\Sigma_\mathrm{r} + \Sigma_\mathrm{g} - 2\sqrt{\Sigma_\mathrm{r}\Sigma_\mathrm{g}}\right)\# \tag{4}$$

Where μ_r and μ_g are the means, and Σ_r and Σ_g are the covariances of the real and generated image features, respectively.

Figure 4 Visual Comparison with Different State-of-the-art Methods in the GTA5-to-Cityscapes Translation Experiment. From Left to Right, the Sequence Displays the Original Image, Followed by the Results of CycleGAN, SRC, TSIT, AttentionGAN, CUT, and our Proposed Method

4.3 Implementation Details

In this study, we utilized the PyTorch [31] deep learning framework to train the network on a single Nvidia A100 GPU (80G). The models were trained with a batch size of 16 and a learning rate of 0.0002. We employed the Adam optimizer, known for its effective learning rate adaptation, to enhance convergence and stability. The training was conducted over 200 epochs to ensure thorough learning and model refinement. The weight of the proposed HOG loss (λHOG) was set to 10, balancing the contribution of HOG features in the loss function and improving feature representation by emphasizing gradient-based features. When computing the HOG features of the image, we set the window size to 256 by 256 pixels, the block size to 6 by 6 pixels, the block stride to 2 by 2 pixels, the cell size to 2 by 2 pixels, and the number of bins to 9. These settings were chosen to ensure a clear and accurate representation of the HOG features. All input images were resized to 256 × 256 pixels to ensure consistency across the dataset, standardizing the input dimensions for uniform training and evaluation. In alignment with the experimental setup for photorealism, we generated images in the Cityscapes style from GTA5 game images. We divided 2,000 images from each dataset as the training set and 500 images as the test set.

4.4 Comparison with State-of-the-art Methods

Table 1 Statistical Comparison with State-of-the-art Methods

	IS	KID	FID
GTA5	-	0.1350	153.0678
CycleGAN	2.6099	0.0813	116.2996
SRC	2.7234	0.0646	98.9844
TSIT	2.6935	0.0477	88.7268
Attention GAN	2.7617	0.0411	82.2881
CUT	2.3735	0.0321	69.2619
Ours	2.6870	0.0310	65.7156

To validate the effectiveness of our proposed method, we performed a comparative analysis using several widely adopted networks for image-to-image translation tasks, including CycleGAN [5], SRC [32], TSIT [33], Attention GAN [7] and CUT [10]. These methods are open-source and have been extensively validated.
As shown in Table 1, our method demonstrates superior performance in terms of the IS metric, which assesses the clarity and diversity of the generated images. Additionally, our method achieved results of 0.0310 for the KID metric and 65.7156 for the FID metric, both of which measure the distance between generated images and the real domain. These results surpass those of all other methods, demonstrating that our approach excels in the task of photorealism compared to other image-to-image translation methods.

4.5 The Visualization Results

As shown in Figure 4, GTA5 represents the original image (a game screenshot), while CycleGAN, SRC, TSIT, Attention GAN, CUT, and Ours represent the images generated by each respective method. It is evident that our method outperforms the others in terms of detail accuracy and correctness in the generated targets. Our method avoids the screen tearing or color distortion that appears in images produced by other methods. Notably, in some methods, the front of the car either disappears or is transformed to blend with the ground, and in the sky, some methods generate vegetation instead. Our method, however, does not exhibit these critical errors, resulting in significantly higher quality images.

5 CONCLUSION

In this paper, we propose an image-to-image translation method based on HOG feature loss, achieving state-of-the-art results in experiments aimed at making GTA5 images appear more realistic within the Cityscapes dataset. Evaluation metrics and visual results demonstrate that our method not only enhances the quality of generated images but also significantly reduces the artifacts commonly seen in other approaches. Moreover, our method exhibits robustness across various challenging scenarios, highlighting its potential for broader applications in image translation tasks. The reduction in visual artifacts not only improves the aesthetic quality but also ensures greater consistency with real-world data. Future work could explore extending this method to a wider range of image-to-image translation tasks, while further enhancing the quality and accuracy of the generated images.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Richardson E, Alaluf Y, Patashnik O, et al. Encoding in style: a stylegan encoder for image-to-image translation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 2287-2296.

[2] Kiani L, Saeed M, Nezamabadi-pour H. Image colorization using generative adversarial networks and transfer learning. 2020 International Conference on Machine Vision and Image Processing (MVIP). IEEE, 2020: 1-6.

[3] Li X, Guo X. SPN2D-GAN: semantic prior based night-to-day image-to-image translation. IEEE Transactions on Multimedia, 2022, 25: 7621-7634.

[4] Isola P, Zhu J Y, Zhou T, Efros A A. Image-to-Image Translation with Conditional Adversarial Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 5967-5976.

[5] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2223-2232.

[6] Cherian A, Sullivan A. Sem-GAN: Semantically-consistent image-to-image translation. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019: 1797-1806.

[7] Tang H, Liu H, Xu D, et al. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. IEEE Transactions on Neural Networks and Learning Systems, 2021, 34(4): 1972-1987.

[8] Emami H, Aliabadi M M, Dong M, et al. SPA-GAN: Spatial attention GAN for image-to-image translation. IEEE Transactions on Multimedia, 2020, 23: 391-401.

[9] Liu M Y, Breuel T, Kautz J. Unsupervised image-to-image translation networks. Advances in Neural Information Processing Systems, 2017, 30.

[10] Han J, Shoeiby M, Petersson L, et al. Dual contrastive learning for unsupervised image-to-image translation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 746-755.

[11] Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005, 1: 886-893.

[12] Richter S R, Vineet V, Roth S, Koltun V. Playing for data: Ground truth from computer games. Proceedings of the European Conference on Computer Vision (ECCV), 2016.

[13] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 3213-3223.

[14] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Advances in Neural Information Processing Systems, 2014, 27.

[15] Radford A. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

[16] Wang T C, Liu M Y, Zhu J Y, et al. High-resolution image synthesis and semantic manipulation with conditional GANs. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8798-8807.

[17] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems, 2017, 30.

[18] Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957, 2018.

[19] Zhang B, Gu S, Zhang B, et al. Styleswin: Transformer-based GAN for high-resolution image generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 11304-11314.

[20] Wyawahare M, Ekbote N, Pimperkhede S, et al. Conversion of Satellite Images to Google Maps Using GAN. International Conference on Innovations in Computational Intelligence and Computer Vision. Singapore: Springer Nature Singapore, 2022: 103-117.

[21] Kim T, Cha M, Kim H, et al. Learning to discover cross-domain relations with generative adversarial networks. Proceedings of the International Conference on Machine Learning. PMLR, 2017: 1857-1865.

[22] Yi Z, Zhang H, Tan P, et al. DualGAN: Unsupervised dual learning for image-to-image translation. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2849-2857.

[23] Huang X, Liu M Y, Belongie S, et al. Multimodal unsupervised image-to-image translation. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 172-189.

[24] Choi Y, Choi M, Kim M, et al. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8789-8797.

[25] Richter S R, AlHaija H A, Koltun V. Enhancing photorealism enhancement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 1700-1715.

[26] Gonzalez-Garcia A, Van De Weijer J, Bengio Y. Image-to-image translation for cross-domain disentanglement. Advances in Neural Information Processing Systems, 2018, 31.

[27] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs. Advances in Neural Information Processing Systems, 2016, 29.

[28] Bińkowski M, Sutherland D J, Arbel M, et al. Demystifying MMD GANs. arXiv preprint arXiv:1801.01401, 2018.

[29] Heusel M, Ramsauer H, Unterthiner T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems, 2017, 30.

[30] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2818-2826.
[31] Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 2019, 32: 8024-8035.
[32] Jung C, Kwon G, Ye J C. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 18260-18269.
[33] Jiang L, Zhang C, Huang M, et al. TSIT: A simple and versatile framework for image-to-image translation. Proceedings of the European Conference on Computer Vision (ECCV), 2020: 206-222.