A Novel Image-to-Image Model: MSF-CycleGAN

Yujing Wang

College of Science, Mathematics and Technology, Whenzhou-Kean University, Wenzhou, China
wangyuji@kean.edu

Abstract. Unpaired image-to-image transformations are often affected by structural distortions and semantic inconsistencies because they rely on pixel-level cyclic consistency constraints. To address these limitations, the paper proposes a multi-scale feature consistency cyclic generative adversarial network, which introduces cyclic constraints in the feature space of the pre-trained Visual Geometry Group (VGG) network. By imposing constraints simultaneously at the low-level structural edges, intermediate textures, and high-level semantic representations, the proposed method enhances the retention of structure while reducing artifacts and background shifts during the transformation process. Additionally, an adaptive weighting mechanism is introduced to automatically balance the influence of different feature scales, enabling the model to capture finer textures while maintaining the integrity of the overall semantic layout. Experimental results on the horse2zebra dataset show that compared to the standard Cycle-Consistent Generative Adversarial Network (CycleGAN), the proposed framework generates clearer stripe textures, more coherent body contours, and better semantic consistency. These advancements demonstrate the robustness of multi-scale feature guidance to stabilize unpaired translation. Going forward, this method offers a promising basis to combine domain-adaptive perceptual features with more computationally efficient feature-level constraints and to evolve feature-level constraints for applications of unpaired translation into a wider application in complex real-world scenarios.

Keywords: Multi-scale feature consistency; Unpaired image-to-image translation; CycleGAN; Feature-level cycle constraint; VGG perceptual features.

1 Introduction

Image-to-image translation is an essential area of research in computer vision, with diverse use cases in image enhancement, style transfer, medical image analysis, and remote sensing image processing [1]. Then most of the classical image translation frameworks depend on paired data, e.g., pix2pix [2]; however, it is often difficult to obtain accurate paired samples in practical scenarios. To solve this problem, Cycle-Consistent Generative Adversarial Network (CycleGAN) proposed an unsupervised approach to image translation based on cycle consistency, which employs bidirectional

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_97

mapping of two domains and successfully translates the image for unpaired data [3]. It has also been applied widely in many fields.

Although CycleGAN has achieved remarkable results, the core pixel-level cycle consistency constraint of CycleGAN has obvious limitations. Firstly, the pixel-level constraints are too strict, requiring the reconstructed image to be nearly consistent in the pixel space. This will cause the generator to tend to produce blurry results, thereby weakening the details of the translated image. Secondly, pixel differences cannot effectively reflect the high-level semantic structure relationship of the image. This leads to problems such as shape distortion, loss of key edges, and inconsistent internal consistency in complex tasks, such as structure-sensitive image translation tasks. Therefore, enhancing the structure-preserving ability of CycleGAN and improving the generation quality still remain a direction worthy of research.

To address the aforementioned issues, this study proposes a multi-scale feature cyclic-consistency CycleGAN. Unlike the traditional CycleGAN framework, the proposed method does not rely solely on pixel-level cyclic constraints. Instead, multi-scale features extracted from different layers of a pre-trained VGG network are utilized to establish cyclic consistency in the feature space [4,5]. Low-level features primarily capture edge and structural information, mid-level features describe texture patterns, and high-level features encode semantic representations. By jointly constraining these multi-level feature layers, the method preserves structural details more effectively and produces clearer and semantically consistent translation results. In addition, an adaptive feature-weight learning mechanism is introduced, enabling the model to automatically adjust the relative importance of different feature scales according to task requirements, thereby improving the flexibility and generalization capability of the framework.

2 Related Work

2.1 Unpaired Image Translation

Unsupervised image translation aims to learn cross-domain mapping without paired training samples. CycleGAN is a representative work in this field, achieving unsupervised style transfer and image transformation by proposing cycle consistency loss. Subsequently, many methods have extended CycleGAN, such as DualGAN and DiscoGAN, which also utilize bidirectional mapping and reconstruction constraints to achieve unsupervised translation [3,6,7]. However, due to the pixel-level cycle consistency that requires pixel-by-pixel similarity for the reconstructed image, these methods still tend to have problems such as structural distortion and texture loss in complex-structured and significantly different-domain tasks.

2.2 Feature Space Constraints

To overcome the limitations of pixel-level loss, feature space loss (feature loss / perceptual loss) has been widely applied in generation tasks. Gatys et al. first demonstrated through neural style transfer that different layers of deep convolutional networks can capture image features at various levels such as edges, textures, and semantic structures [5]. Johnson et al. further proposed to use VGG feature maps as

perceptual loss for image reconstruction and style conversion tasks, significantly improving the perceptual quality [8]. These studies show that the differences in the feature space are more capable of reflecting the image structure than pixel differences, thus providing a theoretical basis for subsequently using multi-scale features to constrain CycleGAN.

2.3 Muti-Scale Feature Representation

Multiscale features play a crucial role in image generation and translation tasks. The VGG network inherently has a hierarchical structure, with different convolutional layers capable of capturing low-level edge features, mid-level texture features, and high-level semantic features [4]. Therefore, it has become the backbone for a large number of visual tasks for feature extraction. Additionally, many improved GAN methods utilize multiscale discriminators, multiscale losses, or pyramid feature structures to enhance the generation quality, such as MSG-GAN [9] and Pix2pixHD [10]. However, these methods mainly focus on the multiscale structure of the generator and discriminator, and rarely directly improve the reconstruction constraints of CycleGAN from the perspective of feature consistency.

2.4 Structure-Preserving Translation

In recent years, some approaches have attempted to enhance structural consistency to alleviate the deformation problem of CycleGAN. For instance, CUT (Contrastive Unpaired Translation) maintains local semantic consistency by introducing contrastive learning; AttentionGAN highlights key regions through an attention mechanism; methods such as SimDAN / GCAN incorporate geometric structure constraints to avoid shape distortion [11,12]. These methods have demonstrated that strengthening constraints at the structural level can significantly improve the performance of unsupervised image translation. However, they do not approach from the perspective of multi-scale semantic structure (edge–texture–semantic) consistency, and there are significant differences from the multi-scale feature cycle consistency method proposed in this paper.

3 Methodology

3.1 CycleGAN baseline

CycleGAN is one of the most representative methods in unsupervised image translation tasks. It achieves image transformation without paired samples by learning bidirectional mappings between two domains. Specifically, CycleGAN consists of two generators G_{AB} and G_{BA}, which are responsible for mapping the source domain A to the target domain B, and mapping the target domain B back to the source domain A. Additionally, the model is equipped with two discriminators D_A and D_B to determine whether the images come from the real data distribution.

The core idea of CycleGN is to introduce the cycle consistency loss. Due to the lack of paired samples, the model cannot directly supervise the translation quality.

Therefore, CycleGAN assumes that the cross-domain mapping should be invertible. To achieve this, CycleGAN defines a cycle consistency loss in the pixel space:

$$\mathcal{L}_{cyc} = |G_{BA}(G_{AB}(A) - A)|_1 + |G_{AB}(G_{BA}(B) - B)|_1 \quad (1)$$

This loss constraint enables the generator to reconstruct the input image after cross-domain transformation, thereby avoiding the mode collapse problem in unsupervised training. Moreover, CycleGAN also keeps adversarial loss (GAN loss), which makes the generated images closer to the distribution of the target domain in appearance:

$$L_{GAN(G_{AB},D_B)} = E_{B\sim p(B)}[logD_B(B)] + E_{A\sim p(A)}\left[log\left(1 - D_B\left(G_{BA}(A)\right)\right)\right] \quad (2)$$

The optimization goal of the generator is to simultaneously minimize the cycle consistency loss and the adversarial loss, while the discriminator maximizes the adversarial loss(min-max loss). By combining the two, CycleGAN achieves satisfactory image translation results even in the absence of paired samples.

However, the pixel-level cycle consistency loss essentially requires that the generator reconstruct each pixel of the image to be in strict correspondence with the input. This excessive constraint often leads to problems such as blurry textures, inconsistent structures, and loss of high-frequency details in the generated images. Especially in tasks where the source domain and the target domain have significant structural or semantic differences, the pixel difference constraint of CycleGAN cannot effectively represent the high-level semantic structure of the images, thereby limiting the quality of unsupervised image translation. Based on this observation, this study further introduced a multi-scale feature cycle consistency mechanism on the basis of the CycleGAN framework to enhance the generator's expression ability at the structural level and improve the detail and semantic consistency of the translated images.

3.2 Multi-Scale Feature Extractor (VGG 19)

In the image translation task, the differences in the pixel space often fail to fully reflect the structural and semantic information of the image. Especially in complex domain transformation scenarios, relying solely on the pixel-level cycle consistency loss of CycleGAN is insufficient to ensure that the generator maintains high-frequency textures, structural edges, and semantic layouts. Therefore, after the generator's output, it is key point that introducing a multi-scale feature extraction mechanism to obtain multi-level structural features from different layers of the deep network.

Specifically, this study employs a pre-trained VGG network as the fixed feature extractor. The different convolutional layers of VGG have a natural hierarchical structure: the shallow features (such as $conv_1$ and $conv_2$) mainly encode local edges and textures, the middle-level features ($conv_3$) contain higher-level regional structures, and the deep features ($conv_4$ and $conv_5$) can capture global semantic relationships. This hierarchical feature framework provides a reliable multi-scale representation for generating the structural consistency of the image.

Set input image as x, generator as G(x), the l-th layer of the feature extractor as $\phi(.)$. Extracting feature maps from multiple convolutional layers can be described by the following formular:

$$F_l(x) = \phi_l(x), F_l\left(G(x)\right) = \phi_l\left(G(x)\right) \quad (3)$$

These features encode the low-level texture structure and high-level semantic information at different scales respectively. By comparing the differences between the

generated image and the input image in the multi-scale feature space, it is possible to effectively reduce structural distortion and improve the texture detail fidelity of the generated image. These multi-layer convolutional features do not require any additional training costs. They merely serve as fixed structure constraints for the feature space, thereby enhancing the structural preservation ability of CycleGAN without increasing the complexity of the model.

3.3 Multi-Scale Feature Cycle Consistency Loss

The Multi-Scale Feature Cycle Consistency Loss assesses how different the feature spaces in the input images are and the cyclic reconstructions produced among different convolutional layers and therefore allows the generator to maintain consistent texture structures and semantic arrangements at different scales.

Let $\phi_\ell(\cdot)$ be the feature extractor in the ℓ-th layer of the pre-trained VGG network. The input image is x, and after being processed by the generator and then back through the network, it is cyclically reconstructed as $\hat{x}$. The loss in feature consistency of layer l is expressed as:

$$\mathcal{L}_{feat}^{\ell}(x, \hat{x}) = |\phi_\ell(x) - \phi_\ell(\hat{x})|_1 \tag{4}$$

This loss is calculated in the feature space, which can capture structural differences that are difficult to reflect by pixel loss, such as edge contours, regional textures, and deep semantic relationships.

To integrate multi-level feature information, the paper extract features from multiple levels, such as the shallow level (conv1_, conv2_), the middle level (conv3_), and the deep level (conv4_). The final multi-scale feature cyclic consistency loss is defined as:

$$\mathcal{L}_{feat-cyc} = \sum \lambda_\ell |\phi_\ell(x) - \phi_\ell(\hat{x})|_1 \tag{5}$$

where, L represents the selected set of feature layers, and λ_ℓ are the weight coefficients for different levels, which are used to balance the contribution of features at different scales to the training process. Compared with the traditional pixel-level cycle consistency loss, the proposed multi-scale feature cycle consistency loss has stronger constraining ability in terms of structural representation.

3.4 Adaptive Weighting mechanism

In addition to using the normalized form of feature differences, this study further introduces an adaptive weight mechanism based on the Softmax function to more stably adjust the contributions of multi-scale features during the training process. Compared to linear normalization, the Softmax weights can provide a smoother and more stable proportion allocation, and naturally possess the characteristics of highlighting large differences and suppressing small differences.

First, define the feature differences of the ℓ-th layer:

$$d_\ell = |\phi_\ell(x) - \phi_\ell(\hat{x})|_1 \tag{6}$$

To obtain the adaptive weights for each layer, this study inputs the difference values into the Softmax function:

$$\lambda_\ell = \frac{e^{d_\ell}}{\sum e^{d_k}} \tag{7}$$

where k belongs to L.

The feature cycle consistency loss based on Softmax is expressed as:

$$\mathcal{L}_{feat-cyc} = \sum_{\ell \in L} \lambda_\ell d_\ell \quad (8)$$

3.5 Overall Objective

To achieve high-quality unsupervised image translation, this study combines adversarial loss, pixel-level cycle consistency loss, and the proposed multi-scale feature cycle consistency loss, and adopts an adaptive weighting mechanism to enhance the effectiveness of feature constraints at different scales. By integrating the above losses, this study constructs the total objective function of the model.

The final optimization objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{GAN}(G_{AB}, D_B) + \mathcal{L}_{GAN}(G_{BA}, D_A) + \alpha \mathcal{L}_{cyc} + \beta \mathcal{L}_{feat-cyc} + \gamma \mathcal{L}_{id} \quad (9)$$

where α, β, γ are the weight factors for the cyclic consistency loss, the multi-scale feature cyclic consistency loss, and the identity loss, respectively. By optimizing the above total loss during the training process, the model can simultaneously maintain the adversarial generation ability, pixel-level consistency, and multi-scale structure-level consistency, thereby effectively improving the overall quality of unsupervised image translation.

4 Experiments

4.1 Dataset

This study employs the most classic horse2zebra unsupervised image translation dataset from the official experiments of CycleGAN. This dataset was first proposed by Zhu et al. in the original paper of CycleGAN and is widely used to evaluate the performance and stability of image-to-image translation tasks.

4.2 Implementation Details

The proposed method is implemented based on a standard CycleGAN framework without additional data augmentation or complex preprocessing. The training data are constructed by directly concatenating images from domain A and domain B along the width dimension, forming aligned pairs for model input. No resizing, cropping, or color transformations are applied beyond this concatenation. The generators adopt a ResNet structure with nine residual blocks and Instance Normalization, then the discriminators adopt the PatchGAN design to classify local image patches. All model components are initialized using normal weight initialization.

The networks are trained using the Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and an initial learning rate of 0.0002 based on a linear decay schedule during the second half of training. The batch size is set to one. A history buffer (ImagePool) is used to store previously generated images, to allow the discriminator to be trained with a mixture of current and past samples to ensure stabilization of adversarial learning. Each time, the model feeds the concatenated image pair, applies forward propagation, performs the GAN and cycle-consistency losses, and updates the network parameters. The training

script occasionally saves model checkpoints and visual outputs for monitoring during training.

4.3 Discussion

Qualitative Analysis. As visual comparison demonstrates, the proposed multi-scale feature cyclic consistency model produces translations with clearer structures and more realistic textures than CycleGAN. The method proposed in Figure 1 maintains the horse's body shape by allowing it to remain true to its original body shape, producing zebra stripes that match the corresponding geometry. CycleGAN, on the other hand, can distort body shapes, blur the lines, creating blurred stripe patterns or induce unrealistic shading. Figure 2 shows that the proposed model preserves the spatial layout of the original zebra and creates smooth horse textures while CycleGAN generates shape deformation, color bleeding, or over-saturated patterns. These findings indicate that the multi-scale feature constraint appropriately decreases structural drift, improving fidelity of the synthesized images.

The main benefit is that feature-level cyclic consistency emerges. Multi-scale VGG features consist of edge-level, texture-level, and semantic-level information simultaneously. Making sure to be consistent across these feature layers prevents the generator from changing the structural content of the input image. Consequently, the translated outputs retain body shape, pose, and scene geometry better than the CycleGAN model which only uses pixel-level cyclic loss. The adaptive weighting mechanism also adjusts the contribution of different feature scales autonomously, resulting in a better model able to cope with high-frequency artifacts, more realistic textures, and smoother appearance in both translation directions.

Fig. 1. Comparison from horse to zebra (Picture credit: Original)

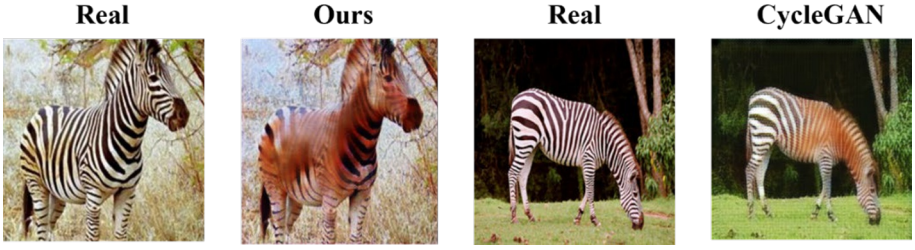

Fig. 2. Comparison from zebra to horse (Picture credit: Original)

Limitation and Future Work. Although this proposed multi-scale and multi-level feature consistency can better maintain the structural integrity, these techniques use features based on VGG at higher depths, and these features cannot fully match all target domains. This dependency may deteriorate the effect when there is a significant difference between domain statistics and natural images. Moreover, the multi-scale feature loss brings more computational overhead compared to the standard CycleGAN. Finally, since this task is asymmetric image transformation, quantitative evaluation is still limited, and the model may still exhibit slight color changes or minor structural defects. Further research can consider using domain-adaptive or task-specific feature extractors to replace VGG and better adapt to various image domains. Smaller feature-based constraints or attention mechanisms can be integrated to reduce computational load while improving the overall structural fidelity. Additionally, incorporating self-supervised semantic constraints may help improve the consistency of the transformation in asymmetric scenarios, thereby enhancing reliability in quantitative evaluation.

5 Conclusion

In this study, a multi-scale feature cycle consistency framework was introduced to address the limitations of traditional CycleGAN in image-to-image transformation without paired images. This framework does not merely rely on pixels but utilizes multi-scale VGG features to preserve the structural, texture, and semantic features during the transformation process. The model achieves a more stable mapping by using feature-level cycle consistency and adaptive weighting, reducing the artifacts caused by common cycle phenomena in CycleGAN.

The qualitative comparison of the "horse2zebra" dataset indicates that the translation results obtained using the proposed method are visually clearer, more coherent, better able to retain textures, and have a more natural color distribution. It is worth noting that in complex situations, the integrity of the structure can be more effectively maintained, while CycleGAN often leads to distortion and inconsistent patterns. The results show that the consistency based on features provides a stronger supervisory signal in unpaired translations.

Overall, the proposed model improves visual realism and structural fidelity in a way that does not rely on paired data, which makes it an easy yet useful upgrade over legacy CycleGAN methods. In the future, more sophisticated perceptual features can be investigated, or attention mechanisms can be integrated to improve the semantic alignment.

References

1. 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), 1125–1134. (2017)

2. Henry, J., Natalie, T., & Madsen, D.: Pix2Pix GAN for Image-to-Image Translation. Technical Report. doi:10.13140/RG.2.2.32286.66887. (2021)
3. Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A.: Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2223–2232. (2017)
4. Simonyan, K., & Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations (ICLR). (2015)
5. Gatys, L. A., Ecker, A. S., & Bethge, M.: A Neural Algorithm of Artistic Style. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVR), 2414–2423. (2016)
6. Yi, Z., Zhang, H., Tan, P., & Gong, M.: DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2849–2857. (2017)
7. Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J.: Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. Proceedings of the 34th International Conference on Machine Learning (ICML), 1857–1865. (2017)
8. Johnson, J., Alahi, A., & Fei-Fei, L.: Perceptual Losses for Real-Time Style Transfer and Super-Resolution. European Conference on Computer Vision (ECCV), 694–711. Springer. (2016)
9. Karnewar, A., & Wang, O.: MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7799–7808. (2020)
10. 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8798–8807. (2018)
11. Park, T., Efros, A. A., Zhang, R., & Zhu, J.-Y.: Contrastive Learning for Unpaired Image-to-Image Translation. European Conference on Computer Vision (ECCV), 319–345. Springer. (2020)
12. Mejjati, Y. A., Richardt, C., Tompkin, J., Cosker, D., & Kim, K. I.: Unsupervised Attention-Guided Image-to-Image Translation. Proceedings of the European Conference on Computer Vision (ECCV), 189–204. (2018)