Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/81/20241216

Enhancing Semantic Consistency in Image-to-Image Translation with an Improved CycleGAN Framework

Runze Fan

Jinan university and university of Birmingham Joint Institution, Jinan University, Jinan, China

rxf218@student.bham.ac.uk

Abstract. This paper presents an enhanced Cycle-Consistent Adversarial Networks (CycleGAN) model aimed at preserving semantic consistency during image-to-image translation, with a focus on complex tasks such as autonomous driving and scientific simulations. The study's key contribution is the incorporation of a pre-trained semantic segmentation model to preserve important characteristics during translation, such as license plates, traffic signs, and pedestrian structures. By introducing a semantic consistency loss alongside the traditional cycle-consistency loss, the proposed approach ensures that key features are retained, even in challenging scenes. Extensive experiments conducted on the Cityscapes dataset demonstrate the effectiveness in maintaining both visual fidelity and semantic accuracy, significantly improving upon the traditional CycleGAN. This method proves particularly valuable in domains where precision is essential, such as cross-domain image generation for autonomous systems and medical imaging. Future research will focus on optimizing the model for real-time applications and exploring multi-domain frameworks to further enhance its performance in diverse environments. Overall, this study offers an efficient image style-transfer solution for preserving semantic integrity without sacrificing translation accuracy.

Keywords: Semantic consistency, CycleGAN, image translation, semantic segmentation.

1. Introduction

In a variety of domains, including science simulations, design, and the arts, Image-to-Image Translation (I2I) is essential. Its major goal is to transfer photos between domains while maintaining the necessary content. Notable advances in this area began with Generative Adversarial Networks (GANs), particularly with pixel-to-pixel (Pix2Pix), which relied on paired data [1]. However, obtaining paired datasets is often impractical, leading to the development of unpaired methods like Cycle-Consistent Adversarial Networks (CycleGAN) [2]. CycleGAN introduced the concept of cycle consistency to achieve domain translation without paired datasets [2]. However, despite its success, CycleGAN suffers from challenges, especially in preserving semantic information during translation. Recent efforts have focused on improving the diversity of generated images by exploring one-to-many mappings [3, 4]. This study revisits CycleGAN's framework to address its semantic consistency limitations, leveraging pre-trained semantic segmentation models to ensure that critical objects and features remain intact during translation. This approach enhances the model's utility in applications requiring high precision.

DOI: 10.54254/2755-2721/81/20241216

In the domain of image-to-image translation, Pix2Pix introduced paired translation using Conditional GANs [1], followed by Pix2PixHD for high-resolution images [5]. However, the difficulty of obtaining paired datasets spurred the development of unpaired methods. Gathering or annotating these kinds of datasets is frequently difficult or costly. Utilizing computer gaming software to create lifelike virtual environments is a viable substitute that offers a controllable and affordable option. Virtually infinite training data may be provided by such software, which can also replicate real-world events that are ordinarily hard to witness. Unfortunately, biases are introduced when using data from synthetic domains, which frequently leads to domain shifts that negatively impact downstream activities' performance. Bousmalis et al. explored unsupervised domain adaptation with GANs [6], while Liu et al. proposed shared latent space translation [7]. Taigman et al. developed a method for cross-domain generation without paired data [8]. CycleGAN, proposed by Zhu et al., solved unpaired translation using cycle consistency but struggled with instability and semantic consistency in complex scenes [2]. Recent models like Multimodal Unsupervised Image-to-Image Translation (MUNIT) [3] and Diverse Image-to-Image Translation (DRIT) [9] introduced multimodal translations, while StarGAN handled multi-domain translation [4]. Diffusion models, such as Denoising Diffusion Probabilistic Models (DDPM), also emerged as powerful generative approaches for style transfer [10]. Despite these advancements, maintaining semantic consistency remains a challenge. Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation (U-GAT-IT) and ACL-GAN have sought to enhance performance through more sophisticated networks or by relaxing cycle-consistency constraints. However, these models often trade off consistency for diversity, making them less suitable for tasks requiring precise one-to-one mapping. This study aims to address these issues by preserving semantic structure without sacrificing consistency.

The introduction of semantic consistency loss in style transfer is designed to preserve crucial details—such as digits on license plates, object types, and other important elements—during the transformation process. For instance, in a street scene style transfer, a car should remain recognizable as a car, and key features like the license plate must stay intact, ensuring that while the visual style changes, essential details for tasks like recognition or identification are maintained. This is particularly critical in fields requiring high accuracy in image translation. The objective of this research is to enhance semantic coherence during image translation by incorporating pre-trained semantic segmentation models. This improves the accuracy of translations in complex scenes by ensuring that important features are retained. The approach integrates a semantic loss function into the CycleGAN training process, preserving essential semantic structures without significantly increasing computational demands. This improved method is valuable for applications where precision is key, such as autonomous driving, scientific simulations, and medical imaging, as it balances efficiency and accuracy.

2. Methodology

2.1. Dataset description and preprocessing

The datasets utilized in this study include the Cityscapes and GTA5 datasets, as shown in Figure 1. The Cityscapes dataset is widely used in the computer vision community, particularly for urban scene understanding and autonomous driving applications [11]. It consists of images captured from a car's perspective across various European cities, featuring 5,000 finely annotated images and an additional 20,000 coarsely annotated images. These images cover 30 different object classes, such as vehicles, pedestrians, buildings, and road signs, making it a robust resource for tasks like semantic segmentation and image translation. Similarly, the GTA5 dataset, derived from a virtual urban environment, contains 24,966 annotated images that are synthetically generated but closely mimic real-world urban scenes, providing additional variability and complexity in training models for semantic segmentation tasks.

Bilinear interpolation was used for preprocessing to scale the pictures from both datasets to 1024x1024 pixels to preserve uniformity in input dimensions across the network. The training data was augmented using techniques like random horizontal flipping and cropping to decrease overfitting and boost variety [12]. To ensure that the inputs stay on a same scale for efficient model training,

DOI: 10.54254/2755-2721/81/20241216

normalization was specifically implemented by scaling the pixel values to a standard range, with a mean of 0.5 and a standard deviation of 0.5, at the same time random horizontal flipping was also applied with a probability of 0.5 to simulate different orientations of the same scene. During the training process, input images are randomly cropped to a size of 768x768 pixels. This operation is a common data augmentation technique aimed at increasing data diversity by cropping different regions of the image, while also enhancing the model's robustness to variations in scale and spatial information. By randomly cropping the original images, the model is exposed to different parts of the input image, thereby preventing overfitting to specific areas and improving generalization across unseen data.

2.2. Proposed approach

The proposed approach enhances CycleGAN by integrating a pre-trained semantic segmentation network to improve semantic consistency during image translation, as shown in the Figure 2. This method adds a semantic consistency loss to ensure the translated images retain the original semantic structures. The loss function is guided by a semantic segmentation module, which provides high-level feature maps to inform the translation process.

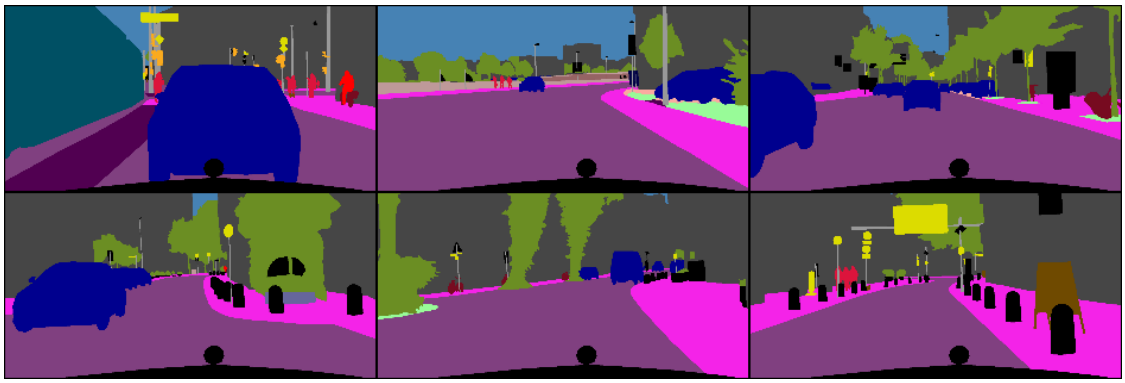

Figure 1. The Cityscape dataset.

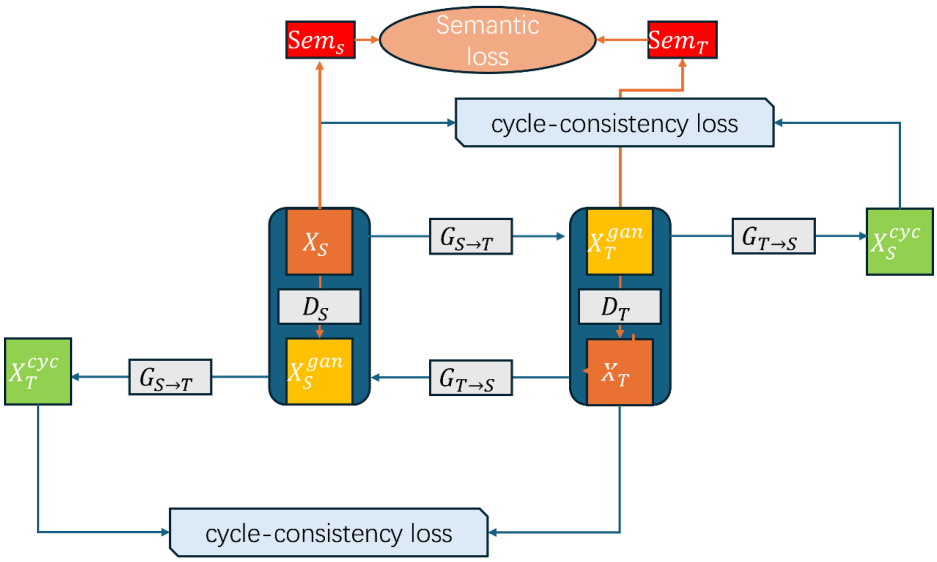

Figure 2. Overview of the improved CycleGAN framework. In addition to the original cycle-consistency loss (blue arrows), which ensures structural integrity after translating between source X_S and target X_T domains, this framework extends the traditional CycleGAN by introducing a semantic loss (orange arrows) to enhance semantic consistency during image translation between domains.

DOI: 10.54254/2755-2721/81/20241216

2.2.1. Generator and discriminator networks. The generator networks $G_{S\to T}$ and $G_{T\to S}$ are central to the CycleGAN architecture, and responsible for translating images between the source and target domains. These generators utilize a U-Net architecture, which is advantageous due to its ability to capture fine-grained details and preserve spatial hierarchies in images [13]. The U-Net structure, consisting of convolutional and deconvolutional layers, is particularly well-suited for tasks that require precise localization of features, such as image translation and segmentation. The discriminator networks D_S andD_T function as PatchGANs, focusing on classifying small patches of an image as real or fake [14]. The patch-based discriminator is effective at concentrating on texture details, which increases the realism of the generated images by allowing the discriminator to focus on localized features. These discriminators provide critical feedback to the generators, encouraging the creation of more realistic and consistent images [15].

2.2.2. Semantic consistency module. The semantic consistency module extracts semantic properties from the source and translated pictures by using a pre-trained semantic segmentation network, such as DeepLabV3 or SegNet. Here, the DRN26 model is used because of its capacity to use dilated convolutions to preserve spatial information at various scales. This aids in the translation process' preservation of semantic structures, which guarantees that the translated picture retains the semantic meaning of the original image. This loss is computed as the L_2-norm between the feature maps SEM_S(source) and SEM_T(translated), guiding the generator to produce translations that retain critical semantic structures of the original image.

2.3. Implementation details

The system is implemented in Python using the PyTorch deep learning framework, selected for its flexibility in constructing custom neural network architectures. The model used in this study is the DRN26 (Dilated Residual Network), which is highly suitable for semantic segmentation tasks due to its ability to maintain spatial information at various levels of depth. The DRN26 architecture incorporates residual connections and dilated convolutions, which enhance feature extraction at multiple scales. Several hyperparameters were fine-tuned to optimize the performance of the model. Several hyperparameters were fine-tuned to optimize the performance of the model. The learning rate for the DRN26 segmentation network was set to 1×10^{-3}, to balance convergence speed and stability. For the CycleGAN architecture, the learning rate for both the generator and discriminator networks was set to 0.0002 after experimenting with various values. A batch size of 2 was used during segmentation training, and 1 for the CycleGAN generators, as memory constraints require small batch sizes for such models. The cycle consistency weight (λ_c) was set to 10, which effectively preserved the source image structure during translation. Meanwhile, the semantic consistency weight (λ_s) was fine-tuned to 1, maintaining a balance between preserving semantic content and not overwhelming the other loss functions. These hyperparameters were selected through empirical testing to ensure optimal model performance.

2.4. Loss function

The proposed approach integrates multiple loss functions to strike a balance between visual realism and semantic accuracy during domain translation. The main loss functions are as follows:

$$L_{GAN}^{S\to T}(G_{S\to T}, D_T, X_T, X_S) = \mathbb{E}_{x_t\in X_T}[logD_T(x_t)] + \mathbb{E}_{x_s\in X_S}\left[\log\left(1 - D_T\left(G_{S\to T}(x_s)\right)\right)\right] \quad (1)$$

$$L_{GAN}^{T\to S}(G_{T\to S}, D_S, X_S, X_T) = \mathbb{E}_{x_s\in X_S}[logD_S(x_s)] + \mathbb{E}_{x_s\in X_T}\left[\log\left(1 - D_S\left(G_{T\to S}(x_t)\right)\right)\right] \quad (2)$$

The GAN loss is fundamental for training the generator to produce realistic images that are indistinguishable from real images in the target domain. Both from the source to the target and from the target to the source, this loss is applied. It ensures the visual fidelity of the translated images, penalizing the generator when the discriminator can distinguish between real and generated images. Equations (1) and (2), for the source-to-target and target-to-source directions, respectively, express this loss function.

DOI: 10.54254/2755-2721/81/20241216

$$L_{cyc}(G_{S\to T}, G_{T\to S}, X_T, X_S) = \mathbb{E}_{x_s\in X_s}[\left|\left|G_{T\to S}\left(G_{S\to T}(x_s)\right) - x_s\right|\right|_1] + \mathbb{E}_{x_t\in X_T}\left[\left|\left|G_{S\to T}\left(G_{T\to S}(x_t)\right) - x_t\right|\right|_1\right] \quad (3)$$

This loss guarantees the ability to reverse when doing the translation process; that is, a picture that is structurally coherent with the original should be produced when translating an image from the source domain to the target domain and back. It is defined as the L_1-norm of the difference between the input image and the image reconstructed after two consecutive translations (equation 3).

$$L_{sem}(Sem_S, Sem_T, G_{S\to T}, G_{T\to S}) = \mathbb{E}_{x_s\in X_s}\mathcal{L}[Sem_S(x_s), Sem_T(G_{S\to T}(x_s))] + \mathbb{E}_{x_t\in X_T}\mathcal{L}[Sem_T(x_t), Sem_S(G_{T\to S}(x_t))] \quad (4)$$

This loss is designed to preserve the semantic content of the source and target images during the translation process. By comparing the semantic features of the input and translated images, the model ensures that critical structures (such as shapes, objects, or important features) remain intact after translation. Semantic features are extracted using segmentation networks Sem_S and Sem_T, which are specifically trained to segment source and target domain images. The semantic consistency loss, defined in equation (4), uses the DRN26 convolutional network to measure discrepancies in the semantic feature space between the original and translated images.

Complete Loss Function: The final loss function (equation 5) combines the above components: the GAN loss from both directions $L_{GAN}^{S\to T}$ and $L_{GAN}^{T\to S}$, the cycle consistency loss L_{cyc}, and the semantic consistency loss L_{sem}. Two hyperparameters, λ_c and λ_s, are introduced to control the relative importance of the cycle consistency loss and semantic consistency loss, respectively. The complete loss function is expressed as:

$$L = L_{GAN}^{S\to T} + L_{GAN}^{T\to S} + \lambda_c L_{cyc} + \lambda_s L_{sem} \quad (5)$$

Here, λ_c ensures that the reconstruction fidelity is maintained, while λ_s emphasizes the preservation of semantic information during the translation. This composite loss framework allows the model to generate visually realistic images while preserving important semantic features, leading to more meaningful and semantically accurate translations compared to traditional CycleGAN models.

3. Result and Discussion

The experimental results of style transfer from GTA5 game screenshots to Cityscapes demonstrate the effectiveness of the model in cross-domain image translation and semantic segmentation tasks, especially the improvement in maintaining semantic consistency, which improves the translation accuracy of the target image. The following is a detailed analysis of the results. As shown in Figure 3 (c) and (d), in the cross-domain conversion experiment from GTA5 game screenshots to cityscapes, the enhanced CycleGAN model performs well in handling complex urban scenes by introducing semantic consistency loss. The model can successfully convert the synthesized GTA5 scenes into realistic cityscape images while retaining important semantic features. In particular, the performance of the model has been significantly improved in the recognition of vehicles, buildings, and road signs.

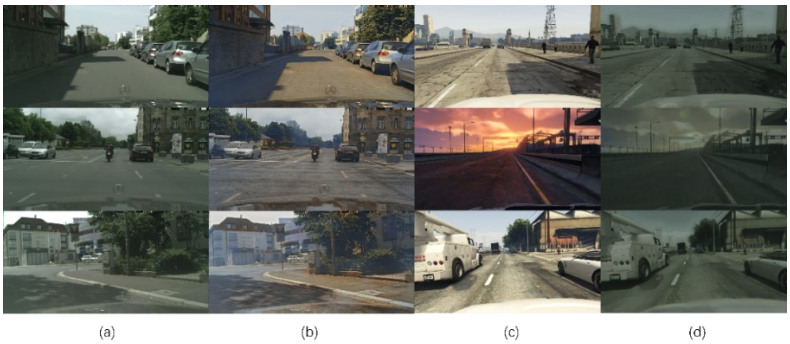

Figure 3. Example images selected from cityscape dataset (a) and GTA5 images dataset(c), along with the translated images (b) and (d), respectively.

DOI: 10.54254/2755-2721/81/20241216

Compared with the source domain model, the performance of the model has been significantly improved in multiple categories, especially in key categories such as "rider" and "vehicle". This improvement is attributed to the fact that the model retains high-level semantic information in the target image during the conversion process, thereby maintaining higher accuracy in the semantic segmentation task.

By observing the experimental results, it can be found that the added semantic consistency loss significantly improves the quality of generated images and the performance of the model in the target domain. The enhanced CycleGAN model can better preserve the details of buildings and roads, making the images it generates in the Cityscapes dataset more consistent with the actual scenes. This is especially important for applications such as autonomous driving that require precise scene understanding. Further experimental results show that the model can better maintain fine-grained semantic information in high-resolution images since the model can effectively maintain the clarity of road markings and license plates while reducing edge blur when processing complex street scenes.

4. Conclusion

In this study, an enhanced CycleGAN model is proposed to improve semantic consistency during image-to-image style transfer, which is particularly important for complex tasks such as autonomous driving and scientific simulations. By combining a pre-trained semantic segmentation model, the proposed method ensures that key features, such as road signs and pedestrians, maintain their semantic structure during translation. The model combines the semantic consistency loss with the standard cycle consistency loss to achieve high-precision translation while preserving essential content. Extensive experiments on the Cityscapes dataset demonstrate the effectiveness of this approach in maintaining visual fidelity and semantic accuracy. The results highlight significant improvements in semantic detail preservation compared to traditional CycleGAN, especially in challenging scenes. Future research aims to improve the model by exploring more advanced segmentation techniques and integrating multi-domain translation frameworks to better handle different urban environments. In addition, optimizing computational efficiency for real-time applications will be a key focus in the future.

References

[1] Isola P Zhu J Y Zhou T Efros A A 2017 Image-to-Image Translation with Conditional Adversarial Networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 1125-1134

[2] Zhu J Y Park T Isola P Efros A A 2017 Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks Proceedings of the IEEE international conference on computer vision pp 2223-2232.

[3] Huang X Liu M Y Belongie S Kautz J 2018 Multimodal Unsupervised Image-to-Image Translation Proceedings of the European conference on computer vision (ECCV) pp 172-189.

[4] Choi Y Choi M Kim M Ha J W Kim S Choo J 2018 StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation Proceedings of the IEEE conference on computer vision and pattern recognition pp 8789-8797.

[5] Wang T C Liu M Y 2018 High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs arXiv preprint arXiv:1711.11585

[6] Bousmalis K Silberman N Dohan D Erhan D Krishnan D 2017 Unsupervised Pixel-Level Domain Adaptation with GANs Proceedings of the IEEE conference on computer vision and pattern recognition pp 3722-3731.

[7] Liu M Y Breuel T Kautz J 2017 Unsupervised Image-to-Image Translation Networks Advances in neural information processing systems p 30.

[8] Taigman Y Polyak A & Wolf L 2017 Unsupervised Cross-Domain Image Generation arXiv preprint arXiv:1611.02200

DOI: 10.54254/2755-2721/81/20241216

[9] Lee H Y Tseng H Y Huang J B Singh M 2018 Diverse Image-to-Image Translation via Disentangled Representations Proceedings of the European conference on computer vision (ECCV) pp 35-51.
[10] Ho J Jain A Abbeel P 2020 Denoising Diffusion Probabilistic Models Advances in neural information processing systems 33 pp 6840-6851.
[11] Cordts M Omran M Ramos S Rehfeld T 2016 The Cityscapes Dataset for Semantic Urban Scene Understanding Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 3213-3223
[12] Shorten C Khoshgoftaar T M 2019 A survey on image data augmentation for deep learning Journal of Big Data 6(1) pp 1-48
[13] Ronneberger O Fischer P Brox T 2015 U-Net: Convolutional Networks for Biomedical Image Segmentation In International Conference on Medical Image Computing and Computer-Assisted Intervention pp 234-241
[14] Li C Wand M 2016 Precomputed real-time texture synthesis with markovian generative adversarial networks In European Conference on Computer Vision pp 702-716
[15] Mirza M Osindero S 2014 Conditional generative adversarial nets arXiv preprint arXiv:1411.1784