IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2017.DOI*

# A Generative Adversarial Network AMS-CycleGAN for Multi-style Image Transformation

**XIAODI RANG[1], ZHENGYU DONG[2], JIACHEN HAN[2], CHAOQING MA[2], GUANGZHI ZHAO[3],AND WENCHAO ZHANG[1]**

[1]School of Information Engineering, Yantai Institute of Technology, Yantai, Shandong 264005, China
[2]School of Computer and Control Engineering, Yantai University, Yantai, Shandong 264005, China
[3]Depatment of Computer Science and Engineering, Jeonbuk National University, jeoniu, Jeollabuk 54896. South Korea

Corresponding author: Xiaodi Rang (e-mail: 1312391660@qq.com).

**ABSTRACT** The objective of image style transfer is to create an image that has the artistic features of a reference style image while also retaining the details of the original content image. Despite the promising outcomes of current approaches, they are still susceptible to generating image information distortion or noise texture problems due to the absence of an effective style representation. As a solution to the aforementioned issues, this paper proposes AMS-CycleGAN (Attention Moment Shortcut-Cycle Generative Adversarial Network), a CycleGAN-based method that achieves style transfer, resulting in artwork that closely resembles hand-painted masterpieces by artists. Initially, the framework makes use of the Positional Normalization-Moment Shortcut (PONO-MS) module, the purpose of which is to retain and transmit structural information in the generator. Additionally, the Multi-Scale-Structural Similarity Index (MS-SSIM) loss is added to strengthen the constraint on the brightness and colour contrast of images. Finally, an attention mechanism module is introduced in the discriminator to emphasize available features and suppress irrelevant features during the style transformation process. According to the experimental results obtained, our method demonstrates a higher level of consistency with human perception when compared to current state-of-the-art methods in image style transfer.

**INDEX TERMS** CycleGAN, Image Style Transfer, Multi-Scale-Structural Similarity Index, Attention Mechanism.

## I. INTRODUCTION

Painting is a captivating and enduring art form that has constantly captured the public's interest. Diversified artworks exhibit a wide array of hues, luminosity, brushstrokes, shapes, and other elements. In the field of computer vision, researchers have been exploring how computer technology can be used to transform ordinary images into artistic paintings. This above process is known as artistic style transfer, which is to extract texture and colour information from the referenced artistic images, and then add this information back to the content image after the transformation. Art style transfer has been utilized across different industries, including film, animation, and gaming, due to its ability to enhance visual effects. Its research has garnered significant attention owing to its value in both scientific and artistic fields.

The latest practices of image style transfer can be categorized into two main categories [1]. The first category is the slow neural method based on online image optimization that generates the stylized image through pixel iteration on the noise image. Based on the distinction of the style loss function, this category can be further categorized into two main types: parameter based methods [2]–[5] and non-parameter based methods [6]–[9]. The limitations of the aforementioned methods include high computational complexity, lengthy processing time, and challenges for real-time applicability. The second group consists of fast neural methods based on offline model optimization. This category encompasses feed-forward network based methods [10]–[15] and generative adversarial network (GAN) [16]–[20] based methods. These approaches leverage pre-training to generate stylized images more efficiently. Specifically, our research focuses on GAN-based methods, where stylized images are generated through an interplay between the generator and the discriminator.

Even though GAN-based methods have made significant

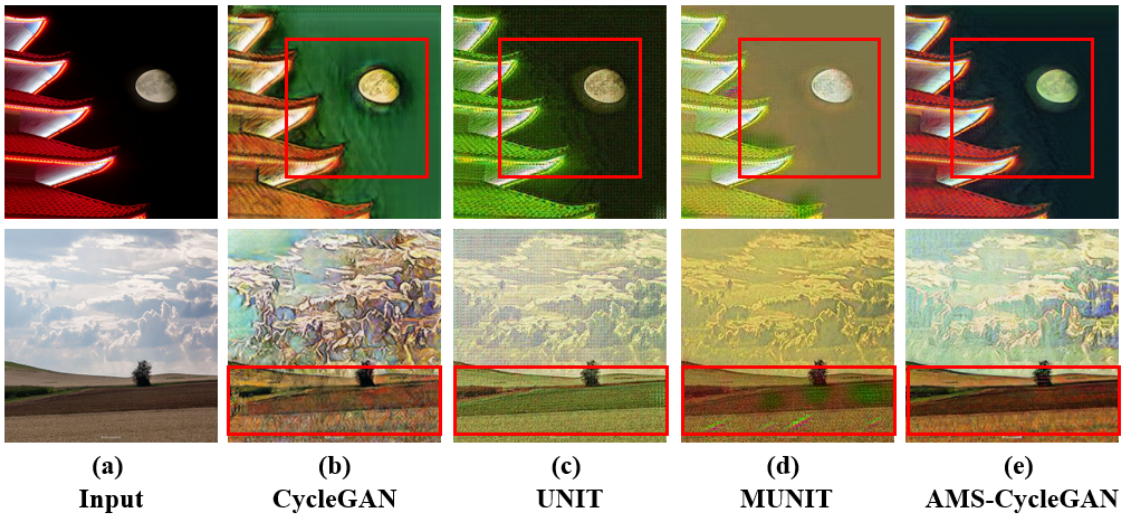


**FIGURE 1. The result of CycleGAN shows the problem of noise texture, UNIT and MUNIT models fail to manifest the distinct characteristics of specific styles and exhibit the problem of noise artifacts in their generation results.**

improvements on unpaired datasets, the generated stylized results are not consistently satisfactory. As shown in Fig.1, the CycleGAN [17], UNIT [21] and MUNIT [19] suffer from issues such as inconsistent semantic information, noisy textures, and colour distortion. To solve this limitation, a structure of style transfer network called AMS-CycleGAN is proposed. The AMS-CycleGAN method is designed to achieve effective style representation and generate high-quality images with improved content consistency and reduced distortion. The results in Fig.1 show that our method can reflect the semantic information from the original photos better. For instance, the color of the sky surrounding the moon and clouds is accurately retained. The contributions of the proposed method are summarized as follows:

1) To enhance the model's ability to capture and utilize the structural features in the images, we have incorporated the Positional Normalization-Moment Shortcut (PONO-MS) module into the generator.
2) To mitigate issues related to brightness, color contrast, and object structure within the generated images, the Multi-Scale-Structural Similarity Index (MS-SSIM) loss combined with L1 norm loss is applied to the reconstructed images.
3) By incorporating an attention mechanism module, the discriminator can recognize the authenticity of the generated image more effectively, and assist the generator to emphasize the key content of the input image.

The following of this paper is organized as: Section II reviews related works about fast neural style transfer; Section III describes the structure of AMS-CycleGAN network; Section IV illustrates the implementation details of the proposed network; Section VI presents and analyzes the comparative experimental results on the public image style transfer datasets; Section VII provides the discussion and the conclusion of this study.

## II. RELATED WORK

The representative works of fast neural style transfer are reviewed and their classification is discussed, including feed-forward network based methods and GAN-based methods. In this section, we will outline the latest and most pertinent research papers.

### A. FEED-FORWARD NETWORK BASED METHODS

The feed-forward network based methods [10], [11], [13], [15], [22]–[24] typically employs a single generator network to generate stylized results, wherein the generator network is trained by minimizing the style loss function to optimize the differences between the generated images and the style images.

Johnson et al. [10] trained feed-forward networks for image transformation tasks by leveraging perceptual losses extracted from deep convolutional neural networks. Ulyanov et al. [11] proposed a texture network that incorporated a multi-scale architecture to learn features from the input image across various dimensions. However, this approach could face challenges when handling intricate textures and styles, potentially resulting in the generation of distorted images. Then, Ulyanov et al. [22] used Instance Normalization (IN) to replace Batch Normalization (BN) [23]. The IN layer performed independent calculations of the mean and variance for each channel and sample, thereby avoiding limitations imposed by batch size. Therefore, Dumoulin et al. [13] proposed Conditional Instance Normalization (CIN) as an extension of IN, in which distinct sets of affine parameters were learned for different styles. This approach assumed that various styles had shared computational dimensions. For instance, many Impressionist painters may have had similar brushstrokes while the choice of colours varied. Adaptive Instance Normalization (AdaIN) was initially introduced by Belongie et al. [15] and by matching the mean and variance of

the content input to those of the style input, and it extended the concept of IN. As the affine parameters in AdaIN were directly computed from the input images, the need for additional training time was eliminated. Afterwards, Park et al. [24] discovered that previous algorithms did not effectively balance global and local style patterns. Therefore, they proposed the Style-Attentional Network (SANet), which learned semantic information between content and style features by spatially rearranging style features based on content features.

In the various normalization schemes of the aforementioned models, a common theme is followed, which involves normalizing across spatial dimensions and discarding extracted statistical data, resulting in a lack of effective style representation. The generated results of these models may also exhibit image information distortion or noise texture issues. Hence, this paper introduces the AMS-CycleGAN model for unsupervised artistic image style transfer, aiming to retain or transmit style feature information in the generated network.

### B. GAN-BASED METHODS

The GAN-based methods train the generator and discriminator through adversarial learning. Following the initial proposal by Goodfellow et al. [16] on GAN-based image style transfer, a considerable number of image-to-image translation models [17], [19]–[21], [25]–[28] based on GANs have been proposed.

Isola et al. [25] proposed the pix2pix method, in which the generator adopted an encoder-decoder structure for image conversion. However, capturing the style of just one or a few images did not adequately capture the full range of an artistic style. To solve this limitation, learning an artist's style from a collection of images became crucial. Therefore, Zhu et al. [17] proposed a network for unsupervised artistic style image transfer and used cycle consistency loss to facilitate the transfer between the source and target domains. This loss function quantified the differences between the mapped images and the original images in the RGB space, thereby preserving the content characteristics of the original images. However, the CycleGAN method frequently encountered challenges, such as the presence of texture noise and colour inconsistency in the generated outcomes. According to Liu et al. [21], it was observed that in the CycleGAN model, input images were mapped onto separate latent spaces. In response to this finding, they proposed the UNIT framework, which assumed a shared latent space such that corresponding images in two domains were mapped to the shared latent space. In addition, Huang et al. [19] proposed the MUNIT model as an extension of the UNIT model, which assumed that the latent space of images could be decomposed into content space and style space. The content encoding encoded the information that should be preserved during the translation process. By sampling different style codes, diverse and multimodal outputs could be achieved in image generation.

Recently, Junho Kim et al. [27] proposed a new method called U-GAT-IT, which incorporated an Adaptive Layer-Instance Normalization (AdaLIN) function and an attention module based on Class Activation Mapping (CAM). This approach enabled the model to prioritize important regions and control the amount of change in shapes and textures within the images, resulting in impressive visual effects in terms of object transformations. Then, a compact network structure NICE-GAN was proposed by Chen et al. [29], which replaced the target domain image encoder by reusing the initial layers of the target domain discriminator. They also proposed a decoupled training paradigm for image conversion, resulting in improved training speed and guaranteed quality of the generated results.

## III. THE PROPOSED METHOD

The proposed network AMS-CycleGAN is based on the CycleGAN [17] model and aims to demonstrate semantic stylization of content images with diverse themes, while effectively retaining the original content information of natural images.

The proposed AMS-CycleGAN comprises two symmetrical Generative Adversarial Networks designed for the source domain $X$ (scenic photographs) and the target domain $Y$ (artist paintings), respectively, as shown in Fig.2. Given training samples $\{x_i\}_{i=1}^{N}(x_i \in X)$ and $\{y_j\}_{j=1}^{M}(y_j \in Y)$ of which the data distribution as $x \sim p_{data(x)}$ and $y \sim p_{data(y)}$ are provided. The generator $G$ is responsible for the transformation from domain $X$ to domain $Y$, with the discriminator $D_y$ assessing the generated images. Conversely, the generator $F$ performs the reverse transformation from domain $Y$ to domain $X$, with the discriminator $D_x$ evaluating the generated images. These operations aim to minimize the significant overlap between the generated images and the target images, while emphasizing the preservation of the original content information of the source images. The discriminators assess the authenticity of the generated images in conjunction with real data, thereby enabling the generators to train towards the intended objective.

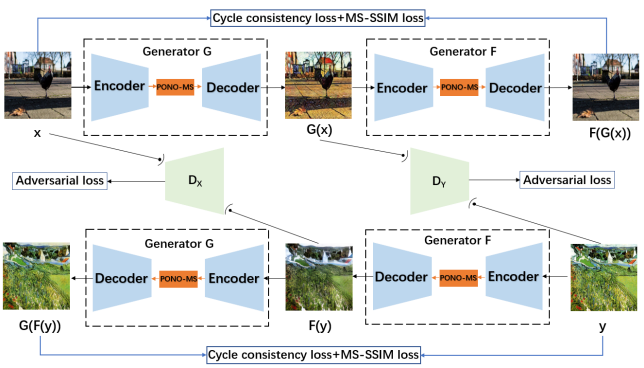


**FIGURE 2. The model architecture of AMS-CycleGAN.**

### A. GENERATOR

The structure of generators $G$ and $F$ are identical in the proposed AMS-CycleGAN model, as shown in Fig.3, with

**TABLE 1. The internal structure of the generator**

| *Part* | *Layer Information* |
|---|---|
| Encoder (Down-sampling) | $\left\{\begin{array}{l} ReflectionPad2d(3) \\ Conv2d(3, 64, k=7, s=1), IN, ReLU \\ Conv2d(64, 128, k=3, s=2, p=1), PONO, IN, ReLu \\ Conv2d(128, 256, k=3, s=2, p=1), PONO, IN, ReLU \end{array}\right\}$ |
| Transformation (Resnet-Block*9) | $\left\{\begin{array}{l} ReflectionPad2d(1) \\ Conv2d(256, 256, k=3), IN, ReLU \\ ReflectionPad2d(1) \\ Conv2d(256, 256, k=3), IN \end{array}\right\}$ |
| Decoder (Up-sampling) | $\left\{\begin{array}{l} MS, ConvTranspose2d(256, 128, k=3, s=2, p=1), IN, ReLU \\ MS, ConvTranspose2d(128, 64, k=3, s=2, p=1), IN, ReLU \\ ReflectionPad2d(3) \\ Conv2d(64, 3, k=7, s=1) \\ Tanh() \end{array}\right\}$ |

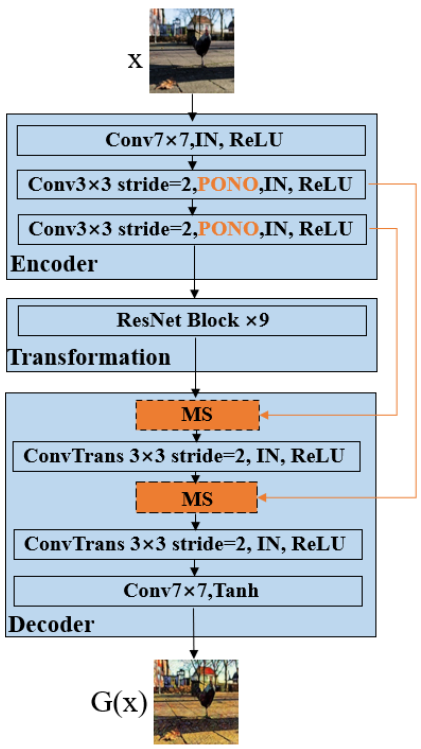


**FIGURE 3. The architecture of AMS-CycleGAN generator. Operations in a block are applied from left to right.**

the internal composition shown in Table.1. The generator consists of three parts: Encoder, Transformation and Decoder. Firstly, the encoder extracts the feature information of the input image $x$, by performing down-sampling using three Conv-IN-ReLU convolutional layers. The input image is mapped to a low-dimensional feature space for data classification and analysis, with the objective of capturing high-dimensional semantic information. Next, inside the transformation module, the target image style is transferred using a 9-layer deep residual network, where each residual network consists of two Conv-IN-ReLU layers with a kernel size of 3. The residual network effectively preserves the content information from source domain images in the generated images, while alleviating the problem of error amplification associated with increasing network depth. Furthermore, in the decoder, two DeConv-IN-ReLU layers are used to restore the low-level features of the input image and collect the semantic information extracted by the encoder. The corresponding feature information is mapped to the pixel locations of the generated image. Finally, the generated image is outputted through a Conv-Tanh layer.

The Encoder and Decoder are connected through a PONO-MS [30] module, which enables the efficient transmission of more pertinent style feature information to the subsequent network layers. The PONO module is located in the down-sampling layers of the Encoder, while the MS module is positioned in the up-sampling layers of the Decoder and receives the output of PONO. As PONO normalizes the channels at fixed pixel locations, it effectively captures the structural information present in the feature maps.

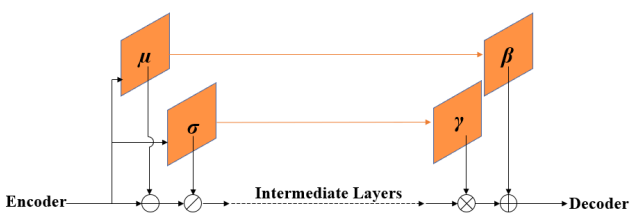


**FIGURE 4. The schematic diagram of the PONO-MS module.**

The schematic diagram of the PONO-MS module is shown in Fig.4. Given an input batch $x \in R^{B\times C\times H\times W}$, the mean ($\mu$) and standard deviation ($\sigma$) information extracted from previous layers are injected directly into the MS layer as the scale ($\gamma$) and displacement ($\beta$) parameters, respectively. Then the calculation of PONO and MS layer are as follows:

$$\mu_{b,h,w}(x) = \frac{1}{C}\sum_{c=1}^{C} x_{b,c,h,w} \tag{1}$$

$$\sigma(x) = \sqrt{\frac{1}{C}\sum_{c=1}^{C}(x_{b,c,h,w} - \mu_{b,h,w}(x))^2 + \varepsilon} \tag{2}$$

$$PONO(x) = \frac{(x - \mu(x))}{\sigma(x)} \tag{3}$$

$$MS(x) = \gamma F(x) + \beta \tag{4}$$

where $c$ is the number of channels, $b$ denotes the batch size, $h$ is the height, and $w$ is the width. In Eq.(2), $\varepsilon$ is small stability constant to avoid divisions by zero and imaginary values due to numerical inaccuracies. $F$ is modeled by the intermediate layers. Also the $\mu$ and $\sigma$ extracted from the input $x$ are directly mapped to $\beta$ and $\gamma$.

### B. DISCRIMINATOR

The discriminator network structure of AMS-CycleGAN is based on the 70*70 PatchGAN [25] model and consists of 6 convolutional layers, as shown in Fig.5. The internal composition is described in Table.2. Firstly, the first to fourth convolutional layers extract feature information from the input image, resulting in 31*31*512 feature maps. Secondly, an attention mechanism module is introduced between the fourth and fifth convolutional layers. By enhancing the interdependence between feature channels, the generator is assisted in selectively focusing on key pixel locations in the image, disregarding or directly filtering out irrelevant parts to obtain the relevant information required for image synthesis. Finally, the output is a matrix $M$ of size $N*N$. Each element $M(i,j)$ of the matrix $M$ corresponds to the receptive field in the original image, and the value of $M(i,j)$ indicates the score of authenticity for the image block in the input image. Additionally, to enhance training stability, spectral normalization (SN) [31] is utilized on the first to fourth convolutional layers of the discriminator.

Let $F \in R^{C\times H\times W}$ represent the intermediate feature map from the fourth layer. Firstly, two distinct spatial feature maps are generated by applying global average pooling and global max pooling operations on $F$. Subsequently, these feature maps are forwarded to a shared network, which comprises a multi-layer perceptron with hidden layers. Following that, weights are calculated along the channel dimension for both the global max pooling feature map and the global average pooling feature map. These weights are applied as activations to the corresponding channels of the intermediate feature map. In other words, the feature maps with weights are concatenated, resulting in a doubling of the number of channels. Finally, a network layer with a 1*1 convolutional kernel is utilized to reduce the channel dimension back to 512. The formula for the shared multi-layer perceptron (MLP) [32] network is as follows:

$$\begin{aligned} M(f) &= concat(F * MLP(GlobalAvgPool(F)), \\ &\quad F * MLP(GlobalMaxPool(F))) \\ &= concat(F * W_1(W_0(F_{gap})), F * W_1(W_0(F_{gmp}))) \end{aligned} \tag{5}$$

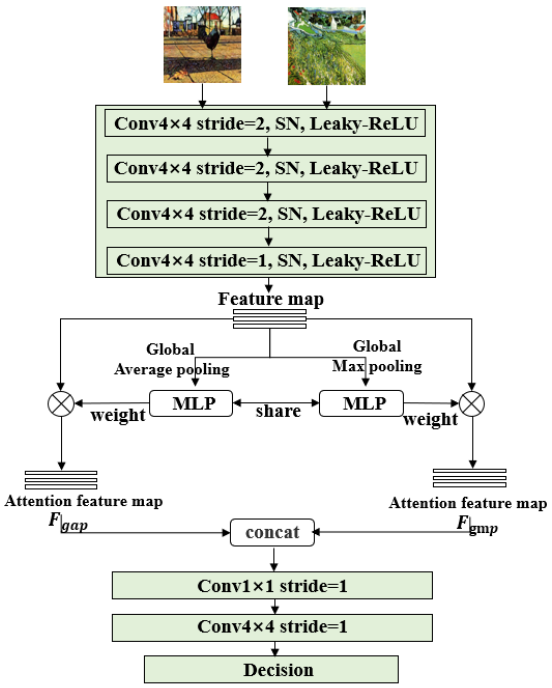


**FIGURE 5. The discriminator of the proposed AMS-CycleGAN.**

where $F_{gap}$ and $F_{gmp}$ represent the one-dimensional feature maps obtained after applying global average pooling and global max pooling operations, respectively. $W_0$ and $W_1$ denote the shared weights of the MLP.

## IV. IMPLEMENTATION DETAILS OF THE EXPERIMENT

### A. LOSS FUNCTION

The loss function consists of four parts: adversarial loss, identity loss, cycle consistency loss [17], and MS-SSIM loss [33].

#### 1) Adversarial Loss

To address the problem of gradient vanishing in traditional GAN, the Least Squares Generative Adversarial Network (LSGAN) [34] approach is introduced, wherein the cross-entropy loss function is replaced by a least squares loss function. LSGAN facilitates stable model training and effectively mitigates mode collapse issues, leading to the generation of more realistic and detailed images.

For the transformation from $X$ to $Y$, the objective function for the generator $G$ and discriminator $D_Y$ is:

$$\begin{aligned} \min_{D_Y} L_{lsgan}(G, D_y, X, Y) = &\frac{1}{2}E_{y\sim data_{(y)}}[(D_Y(y) - 1)^2] \\ &+\frac{1}{2}E_{x\sim data_{(x)}}[(D_Y(G(x)))^2] \end{aligned} \tag{6}$$

$$\min_{G} L_{lsgan}(G, D_y, X) = \frac{1}{2}E_{x\sim data_{(x)}}[(D_Y(G(x)) - 1)^2] \tag{7}$$

**TABLE 2. The internal structure of the discriminator**

| *Part* | *Layer Information* |
|---|---|
| The first to four layers | $\left\{ \begin{array}{c} Conv2d(3,64,k=4,s=2), SN, LeakyReLU \\ Conv2d(64,128,k=4,s=2), SN, LeakyReLU \\ Conv2d(128,256,k=4,s=2), SN, LeakyReLU \\ Conv2d(256,512,k=4,s=1), SN, LeakyReLU \end{array} \right\}$ |
| Attention mechanism module | $\left\{ \begin{array}{c} GlobalAverage, MaxPooling(512,1024) \\ MLP-(N1), MultiplyMLP_{weights} \end{array} \right\}$ |
| The last two layers | $\left\{ \begin{array}{c} Conv2d(1024,512,k=1,s=1), SN, LeakyReLU \\ Conv2d(512,1,k=4,s=1), SN, LeakyReLU \end{array} \right\}$ |

For the transformation from $Y$ to $X$ , the objective function for the generator F and discriminator $D_X$ is:

$$\min_{D_X} L_{lsgan}(F, D_x, Y, X) = \frac{1}{2} E_{x \sim data_{(x)}}[(D_X(x) - 1)^2] + \frac{1}{2} E_{y \sim data_{(y)}}[(D_X(F(y)))^2] \tag{8}$$

$$\min_{F} L_{lsgan}(F, D_x, Y) = \frac{1}{2} E_{y \sim data_{(y)}}[(D_X(F(y)) - 1)^2] \tag{9}$$

Therefore, the adversarial loss for generators ($G$ and $F$) and discriminators ($D_y$ and $D_x$) can be defined as follows:

$$\begin{aligned} \min_{D_{(XY)}} L_{lsgan} = & \frac{1}{2} E_{x \sim data_{(x)}}[(D_X(x) - 1)^2] \\ & + \frac{1}{2} E_{y \sim data_{(y)}}[(D_X(F(y)))^2] \\ & + \frac{1}{2} E_{y \sim data_{(y)}}[(D_Y(y) - 1)^2] \\ & + \frac{1}{2} E_{x \sim data_{(x)}}[(D_Y(G(x)))^2] \end{aligned} \tag{10}$$

$$\begin{aligned} \min_{F,G} L_{lsgan} = & \frac{1}{2} E_{y \sim data_{(y)}}[(D_X(F(y)) - 1)^2] \\ & + \frac{1}{2} E_{x \sim data_{(x)}}[(D_Y(G(x)) - 1)^2] \end{aligned} \tag{11}$$

### 2) Identity Loss

To ensure consistency in color composition between input and output images, the concept of identity loss is incorporated. By using the generator $G$ to generate artistic style images in the target domain $Y$, an image $y$ from domain $Y$ is inputted into the generator $G$, and the generated result should ideally be the same image $y$, that is $G(y) \approx y$. This demonstrates that the generator $G$ is capable of generating artistic style images with the target domain $Y$. The formula for the identity loss is presented as follows:

$$\begin{aligned} L_{identity}(G, F) = & E_{y \sim P_{data_{(y)}}}[\| G(y) - y \|_1] \\ & + E_{x \sim P_{data_{(x)}}}[\| F(x) - x \|_1] \end{aligned} \tag{12}$$

### 3) Cycle Consistency Loss

The cycle consistency loss is applied to the reconstruction of images with the purpose of alleviating mode collapse issues. By utilizing the generator $G$ to convert images $x$ from domain $X$ to domain $Y$, the image $F(G(x))$ generated by the generator F can be transformed back to the original input image $x$. The difference between the reconstructed image and the actual image is computed using the L1 norm [35], as shown in the following equation:

$$\begin{aligned} L_{cycle}(G, F, X, Y) = & E_{y \sim P_{data_{(y)}}}[\| G(F(y)) - y \|_1] \\ & + E_{x \sim P_{data_{(x)}}}[\| F(G(x)) - x \|_1] \end{aligned} \tag{13}$$

### 4) MS-SSIM Loss

To capture the details and structural information of images and enhance the similarity of reconstructed images, the MS-SSIM [33] loss function is introduced into the reconstruction loss. MS-SSIM is an extension of SSIM [36] that incorporates the fusion of images at different resolutions to calibrate the parameters between images of different scales. For two images $x$ and $y$, each with dimensions $H$ and $W$, the means ($\mu_x$, $\mu_y$) and variances ($\sigma_x^2$, $\sigma_y^2$) can be computed using the following formulas:

$$\mu_x = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{(i,j)} \tag{14}$$

$$\mu_y = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} y_{(i,j)} \tag{15}$$

$$\sigma_x^2 = \frac{1}{HW - 1} \sum_{i=1}^{H} \sum_{j=1}^{W} (x_{(i,j)} - \mu_x) \tag{16}$$

$$\sigma_y^2 = \frac{1}{HW - 1} \sum_{i=1}^{H} \sum_{j=1}^{W} (y_{(i,j)} - \mu_y) \tag{17}$$

then the covariance ($\sigma_{xy}$) of the two images is calculated as:

$$\sigma_{xy} = \frac{1}{HW - 1} \sum_{i=1}^{H} \sum_{j=1}^{W} (x_{(i,j)} - \mu_x)(y_{(i,j)} - \mu_y) \tag{18}$$

the formulas for calculating the luminance ($l$), contrast ($c$) and structure ($s$) information between image x and image y are as follows:

$$l_{(x,y)} = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{19}$$

$$c_{(x,y)} = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{20}$$

$$s_{(x,y)} = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{21}$$

and the SSIM loss is calculated as:

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \tag{22}$$

Since MS-SSIM loss refers to the calculation of SSIM values at multiple scales, the formula is expressed as follows:

$$MS\text{-}SSIM(x,y) = [l_M(x,y)]^{\alpha M} \cdot \prod_{j=1}^{M} [c_j(x,y)]^{\beta j} [s_j(x,y)]^{\gamma j} \tag{23}$$

The width and height of the generated images decrease by a factor of $2^{(M-1)}$ at different scales, where the value of M is set to 5 in Eq.(23). The MS-SSIM loss function calculates the contrast and structural similarities of the generated images at dimensions $256 * 256, 128 * 128, 64 * 64, 32 * 32$, and $16 * 16$, while the luminance contrast is only considered at the dimension $16 * 16$. Therefore, the expression for the MS-SSIM loss function is as follows:

$$\begin{aligned} L_{MS\text{-}SSIM}(G,F) = [1 - MS\text{-}SSIM(x, F(G(x))] \\ + [1 - MS\text{-}SSIM(y, G(F(y))] \end{aligned} \tag{24}$$

Taking into account the aforementioned steps, the total loss for training the entire network can be expressed as follows:

$$\begin{aligned} L_{total_{gan}} = \lambda_1 L_{lsgan} + \lambda_2 L_{identity}(G,F) \\ + \lambda_3 L_{cycle}(G,F,X,Y) + \lambda_4 L_{MS\text{-}SSIM}(G,F) \end{aligned} \tag{25}$$

where the parameters $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ are used to control the linear combination of these losses.

### B. DATA PREPARATION AND TRAINING DETAILS

The style transfer capability of AMS-CycleGAN has been demonstrated on two datasets proposed by Zhu et al. [17] in the CycleGAN model: photo2monet and photo2vangogh. The training dataset consists of 6,287 landscape photos, learning from the artistic style images of VanGogh (400 images) and Monet (1,072 images) datasets. To conduct the testing, there were 751 natural photos, 400 style photos of VanGogh, and 121 style photos of Monet. All images were resized to a dimension of 256*256.

During the training stage, the generator and discriminator adopted the Adam optimization algorithm with $(\beta_1, \beta_2) = (0.5, 0.999)$. The batch size was set to 1, and the number of epochs was set to 200. For the first 100 epochs, the initial learning rate was set to 0.0002, and for the remaining 100 epochs, the learning rate was linearly decayed to 0. In the linear combination of the total loss function, in Eq.(25), $\lambda_1$ and $\lambda_4$ were both set to 1.0, and $\lambda_2$ and $\lambda_3$ were set to 5.0 and 10.0, respectively.

The experiments were executed under Ubuntu 18.04.1 system, using Intel (R) Xeon (R) Platinum 8255C with 47 GB of memory and 24GB NVIDIA GeForce RTX 3090 GPU.

## V. THE EXPERIMENTAL RESULTS

### A. QUALITATIVE COMPARISON

To validate the visual quality and style controllability of the proposed network model, AMS-CycleGAN, a qualitative comparison was conducted between AMS-CycleGAN and other models on the photo2monet and photo2vangogh datasets, as well as the monet2photo and vangogh2photo datasets. The models compared include CycleGAN [17], MUNIT [19], UNIT [21], U-GAT-IT [27], and NICE-GAN [29]. CycleGAN [17] and AMS-CycleGAN were implemented using the PyTorch version, with a training epoch set to 200. And MUNIT [19], UNIT [21], U-GAT-IT [27], and NICE-GAN [29] were implemented in the official TensorFlow version with a training iteration of 1,000,000. The results demonstrated that the images generated by the AMS-CycleGAN model effectively preserved the semantic layout of the input images while imitating the specific styles of the target artists.

The comparison of the aforementioned networks in transforming natural photos into artistic images is illustrated in Fig.6 and Fig.7. It can be observed that even under the same style, the stylized images generated by different network models exhibit distinct visual effects. The UNIT [21] model produces images with transitional artifacts and colour distortions. For instance, in the second row of generated images in Fig.6(b), the color of the grass transitions from green to orange. Additionally, in the red range of Fig.7(b), the sky and city colors appear pale in the first and second rows, while the third row images suffer from noise artifacts around the grass. This limitation arises from the UNIT [21] model's assumption of a shared latent space for handling image transformations, necessitating the consideration of the suitability of the latent space and the similarity between image domains. The MUNIT [19] model is an extension of the UNIT [21] model, which improves the quality of generated images to some extent. However, it encounters issues such as the loss of low-level semantic information and the presence of localized noisy textures. For example, in the red-boxed regions in column (c) of Fig.7, the colors of the tree trunk and roof appear as green, which do not correspond to the colors in the original image. The MUNIT [19] model's process of decomposing the input image into content and style codes, where the content code preserves the primary content information while randomly sampling different style codes, results in diverse and multi-modal images.

The CycleGAN [17] model, despite its constraint on the identity mapping loss function, exhibits issues of noise textures and image over-transfer. In Fig.6(d), the first and second rows show noise textures in the sky, while the third row exhibits image over-transfer in the white clouds. Additionally, Fig.7(d) shows blurry boundaries in the second row of modern buildings. These limitations arise from the excessive constraints imposed by the cyclic consistency loss in CycleGAN [17], which directly measures the discrepancy between the stylized output and the content image in the RGB space, resulting in an overly strong constraint on any changes

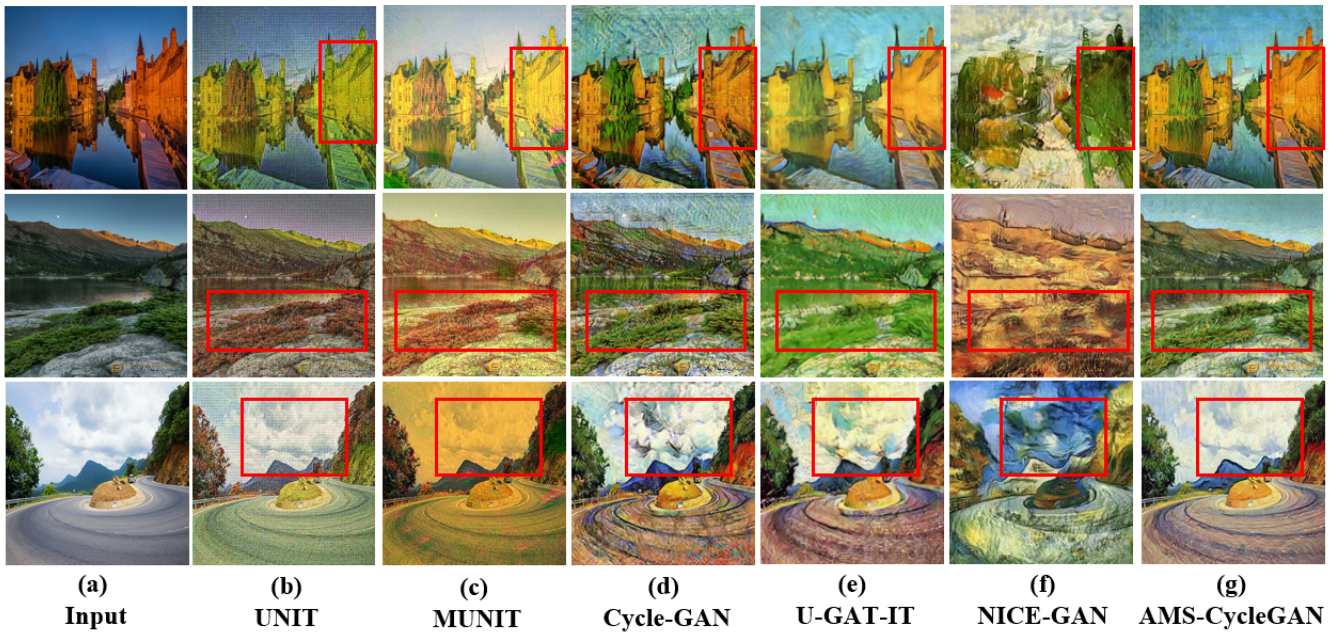


**FIGURE 6.** **Comparison with the mentioned image style transfer methods on photo to VanGogh.**

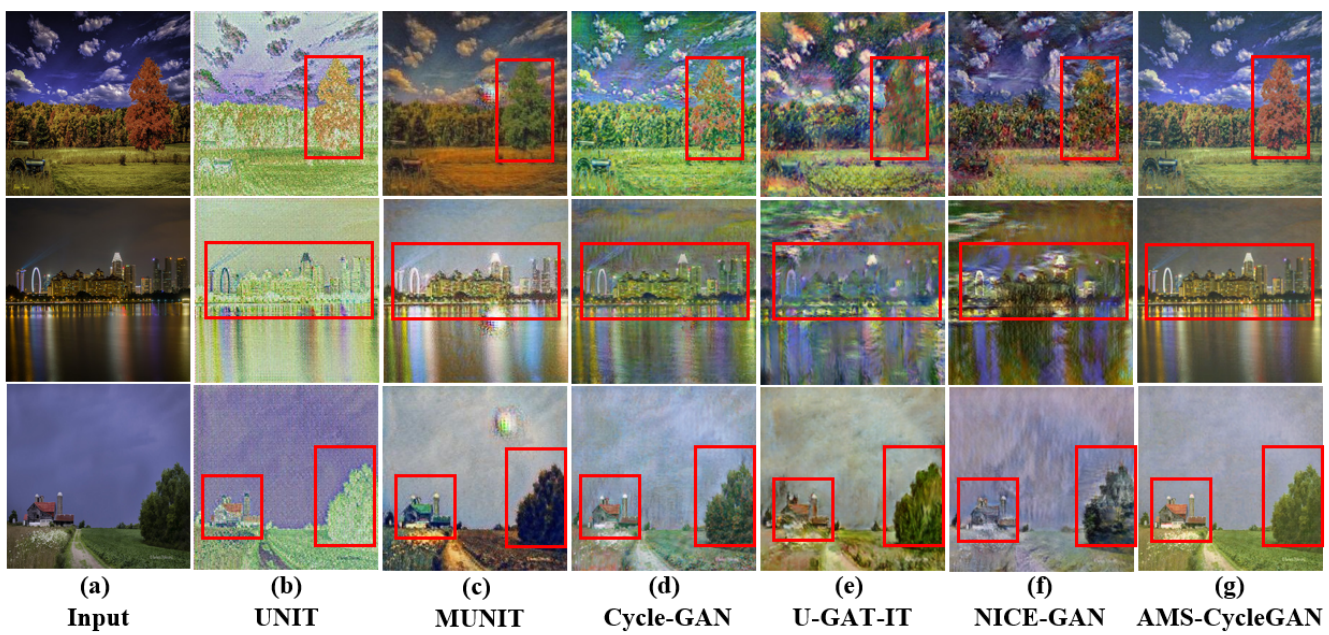


**FIGURE 7.** **Comparison with the mentioned image style transfer methods on photo to Monet.**

in appearance or structure between the input and output. However, it should be noted that the structural differences between content and style images vary depending on the style. Therefore, the influence of cyclic consistency loss on the results varies for different styles.

The U-GAT-IT [27] and NICE-GAN [29] models generate images with issues including noise artifacts, color distortions, and blurred boundaries. These problems can be observed in the surrounding areas of the grass and hills in the second row of Fig.6(e), as well as in the image distortions in the first and second rows of Fig.7(e). Also, in Fig.6 and Fig.7(f), the buildings appear green while the sky appears blue. These limitations can be attributed to the utilization of attention maps in the U-GAT-IT [27] and NICE-GAN [29] models for selecting features, which guide the models to learn relevant parameters from the dataset to control variations in object shapes and image textures. However, these models emphasize object transformations, such as the conversion of a cat to a dog. In contrast, the AMS-CycleGAN model produces superior visual effects by generating retro-styled images that mimic the styles of VanGogh and Monet, while preserving the original content of the input images. This method significantly alleviates the problem of noise textures and colour distortions in the generated images. For instance, as shown in Fig.6 and Fig.7(g), the textures of clouds and the sky in the images are preserved intact. In Fig.7(g), the images in the third row exhibit clear boundaries around the bushes.

## VI. THE EXPERIMENTAL RESULTS

### A. QUALITATIVE COMPARISON

Fig.8 and Fig.9 show the results of the monet2photo and vangogh2photo datasets, which are natural photos converted from artistic images. In the results of the UNIT [21] and MUNIT [19] model conversions, there are significant deviations from the original natural images. The UNIT [21] model's transformed images exhibit problems such as the loss of high-level semantic information, mismatched content compared to the original images, and colour distortion. As shown in Fig.8(b), the pile of haystacks is presented as green in the second row, and the generated image in the third row omits the display of houses. In Fig.9(b), the pale purple sky in the second row is transformed into deep blue, and the visual appearance of the images in the third row presents dull color and low contrast. The transformed images by the MUNIT [19] model, as demonstrated in the first row of Fig.8(c), result in the conversion of bridges into trees, and the transformation of the grass clumps into the ocean in the third row. In Fig.9(c), the shrubs in the second row are transformed into mounds, and the surroundings of the windmill in the third row suffer from the loss of advanced semantic information. The images generated by the CycleGAN [17] model exhibit issues such as low-level semantic information loss and noise textures. In Fig.8(d), the second row shows a conversion where the color of the grassland is transformed into gray.

The U-GAT-IT [27] model's generated images suffer from semantic information loss. In Fig.8(e), the first row exhibits a conversion where the background is transformed into black, and the third row fails to include the flowers and houses present in the input image. Additionally, in Fig.9(e), the generated image in the first row lacks the content information of the lighthouse. The NICE-GAN [29] model generates images with color inconsistencies compared to the original images, such as the bridge in the first row of Fig.8(f) and the sky in the second row of Fig.9(f). The AMS-CycleGAN model preserves the details in the original image as much as possible and retains high-level semantic information. For example, in the third row of Fig.8(g), the color of the red flowers is preserved as red. In Fig.9(g), the background color in the second row matches the pale purple of the original image, and the semantic information of the vine branches is well-preserved in the generated image.

### B. QUANTITATIVE COMPARISON

For quantitative assessment, the Inception Score (IS) [37] and Fréchet Inception Distance (FID) [38] were utilized as evaluation metrics to quantify the image quality.

The Inception Score (IS) is a widely used evaluation metric in image-to-image translation tasks. It leverages a pre-trained image classifier (Inception Network V3 [39]) to evaluate images based on the entropy of their class probability distribution. A higher IS score indicates that the generated images exhibit more diversity and cover a wider range. The definition of IS is:

$$IS = exp(E_{x \sim p_g} D_{KL}(p(y \mid x)) \| p(y))) \tag{26}$$

where $x$ denotes one generated image, and $y$ is the label predicted by the Inception model. $p(y \mid x)$ represents the probability distribution that the picture belongs to each category.

**TABLE 3. Quantitative evaluation and comparison with existing methods in photo2monet and photo2vangogh datasets.**

| | photo-monet | | photo-vangogh | |
|---|---|---|---|---|
| Model | IS | FID | IS | FID |
| UNIT | 5.01±0.43 | 154.39 | 5.18±0.58 | 101.72 |
| MUNIT | 4.73±0.43 | 109.31 | 5.10±0.70 | **97.63** |
| CycleGAN | 5.08±0.63 | 92.83 | 4.18±0.32 | 151.16 |
| U-GAT-IT-light | 3.72±0.23 | 127.85 | 4.33±0.31 | 187.52 |
| NICE-GAN-light | 3.73±0.31 | 117.98 | 3.43±0.19 | 199.70 |
| AMS-CycleGAN | **6.01±0.89** | **74.64** | **5.33±0.76** | 113.28 |

**TABLE 4. Quantitative evaluation and comparison with existing methods in monet2photo and vangogh2photo datasets.**

| | monet-photo | | vangogh-photo | |
|---|---|---|---|---|
| Model | IS | FID | IS | FID |
| UNIT | 3.06±0.27 | 214.97 | 2.85±0.29 | 200.28 |
| MUNIT | 2.00±0.24 | 185.53 | 2.30±0.16 | 229.79 |
| CycleGAN | 3.42±0.37 | 118.35 | 4.36±0.28 | 168.30 |
| U-GAT-IT-light | 2.90±0.31 | 147.99 | 3.89±0.15 | 183.93 |
| NICE-GAN-light | 2.99±0.31 | 144.22 | 3.18±0.26 | 203.61 |
| AMS-CycleGAN | **3.46±0.35** | **112.90** | **4.95±0.49** | **149.68** |

The Fréchet Inception Distance (FID) serves to capture the similarity between real and generated images. The Inception

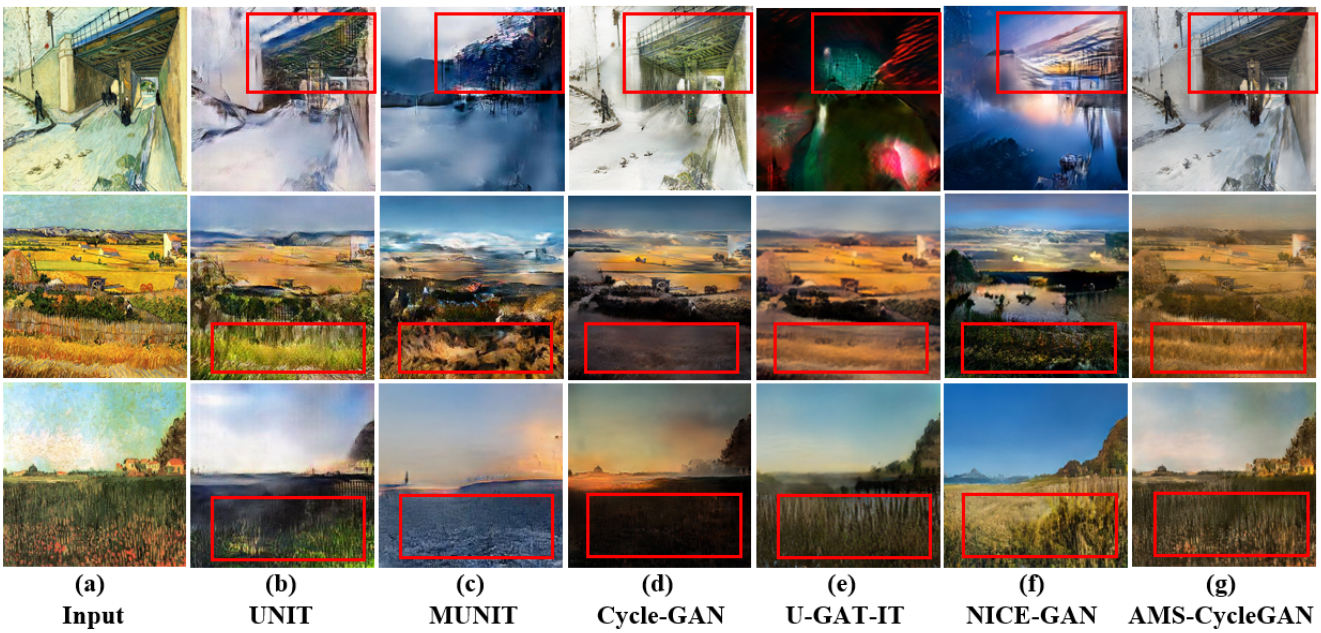


**FIGURE 8.** **Comparison with the mentioned image style transfer methods on VanGogh to photo.**

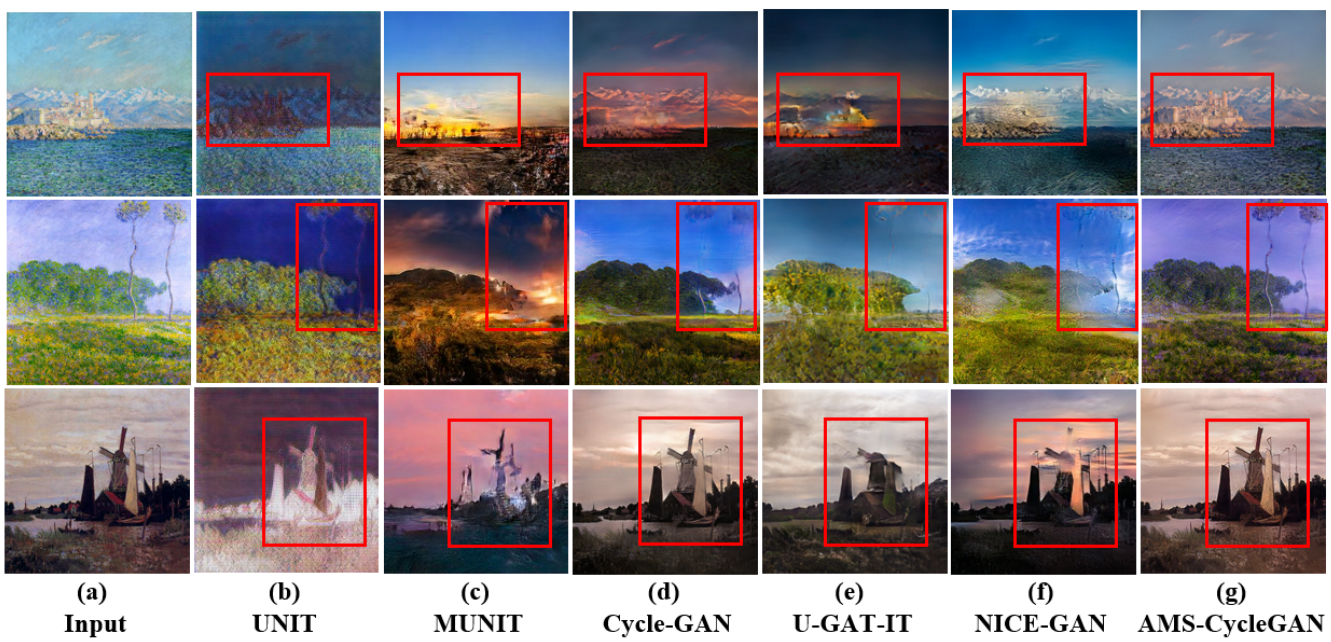


**FIGURE 9.** **Comparison with the mentioned image style transfer methods on Monet to photo.**

Network V3 is utilized as a feature extractor for FID and does not rely on it to determine the specific categories of the images. A lower FID indicates a higher degree of shared visual attributes between the generated images and the original images. FID is defined as follows:

$$FID = \|\mu_{data} - \mu_g\|^2 + tr(\Sigma_{data} + \Sigma_g - 2(\Sigma_{data}\Sigma_g)^{\frac{1}{2}}) \tag{27}$$

where $\mu_{data}$, $\mu_g$, $\Sigma_{data}$, and $\Sigma_g$ are the means and covariances of the samples from the distribution of the source domain and the distribution of the generated data, respectively.

Tables 3 and 4 showcase the numerical results of the UNIT [21], MUNIT [19], CycleGAN [17], U-GAT-IT-light [27], NICE-GAN-light [29], and AMS-CycleGAN [17] models in terms of the IS and FID evaluation metrics. The AMS-CycleGAN model achieves higher IS scores than other models on the photo2vangogh and photo2monet datasets, as well as the vangogh2photo and monet2photo datasets. Furthermore, it also achieves decent results in the FID test scores. The quantitative evaluation results align with the qualitative evaluation results, providing evidence that the inclusion of the PONO-MS module, MS-SSIM loss, and attention module in the AMS-CycleGAN model preserves more content features and effectively showcases the style representation of the images.

### C. ABLATION STUDIES

The previous comparative experiments have validated the effectiveness of the AMS-CycleGAN model in preserving image textures during the mutual transformation of photos and artistic images. In this section, the PONO-MS module, MS-SSIM loss, and attention mechanism module are evaluated for their constraining effects on the generated images. Therefore, comparative experiments are conducted with the addition of the PONO-MS module, the addition of MS-SSIM loss + PONO-MS module, and the AMS-CycleGAN model with all three components (PONO-MS module + MS-SSIM loss + attention mechanism module). Fig.10 and Fig.11 showcase the qualitative results of the ablation experiments.

Fig.10 illustrates the generated results of transforming natural photos into artistic images. In the results that only involve the PONO-MS module, there are issues with excessive color contrast and texture noise, as evidenced by the clouds in Fig.10 (b). When incorporating both the PONO-MS module and MS-SSIM loss, there is a problem with inconsistencies between the generated brushstrokes and the input image. For instance, in the first row of Fig.10 (c), the fur of the fox is transformed into green. AMS-CycleGAN enhances the image clarity of stylized images by effectively mitigating noise problems, as observed around the mountains in Fig.10. Fig.11 demonstrates the generated results of transforming artistic images into natural photos. In the results that solely incorporate the PONO-MS module, issues arise regarding the loss of high-level semantic information, such as faces, fruits, and cups. Likewise, the inclusion of both the PONO-MS module and MS-SSIM loss fails to preserve the complete integrity of objects in the input image. In contrast, AMS-CycleGAN retains the facial information of the characters in the generated images, as demonstrated by the first row of Fig.11 (d).

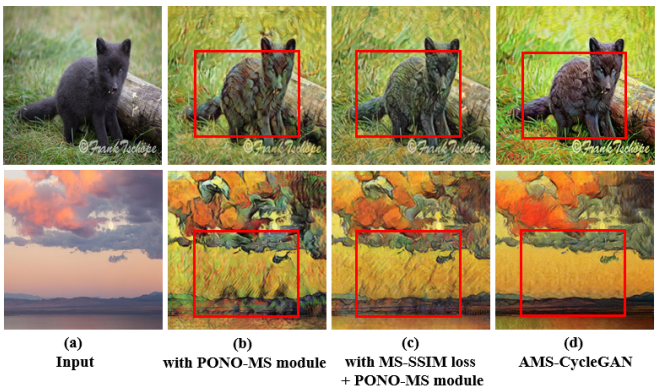


**FIGURE 10.** Qualitative comparisons against ablations of the proposed method. From left to right: (a) content images; (b) with PONO-MS module; (c) with the MS-SSIM loss + PONO-MS module; (d) AMS-CycleGAN (PONO-MS module + MS-SSIM loss + attention mechanism module).

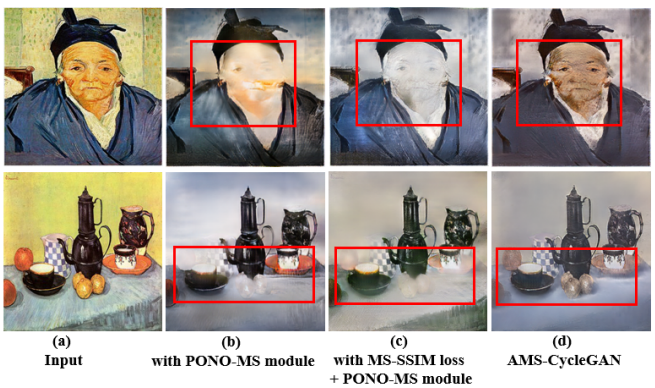


**FIGURE 11.** Qualitative comparisons against ablations of the proposed method. From left to right: (a) artistic images; (b) with PONO-MS module; (c) with the MS-SSIM loss + PONO-MS module; (d) AMS-CycleGAN (PONO-MS module + MS-SSIM loss + attention mechanism module).

**TABLE 5.** Comparison with existing methods on IS and FID.

| | photo-vangogh | | vangogh-photo | |
|---|---|---|---|---|
| Model | IS | FID | IS | FID |
| with PONO-MS module | 4.30±0.34 | 160.92 | 4.23±0.33 | 172.18 |
| with the MS-SSIM loss + PONO-MS module | 4.67±0.51 | 137.45 | 4.44±0.34 | 163.01 |
| AMS-CycleGAN | **5.33±0.76** | **113.28** | **4.95±0.49** | **149.68** |

Table 5 presents the quantitative comparison of the ablation experiments for the AMS-CycleGAN model on the photo2vangogh and vangogh2photo datasets. Compared to other methods, the AMS-CycleGAN model exhibits the highest IS scores and the lowest FID scores, indicating superior image generation capabilities. Considering both qualitative

and quantitative results, the findings can be summarized as follows: Firstly, the PONO-MS module complements the style features in the generated images. Secondly, when only the PONO-MS module + MS-SSIM loss is utilized, there is a noticeable reduction in texture noise issues in the generated images. Thirdly, the attention mechanism assists the generator in emphasizing the crucial content of the input image, enabling the generated images to retain the content information of the input image as much as possible. Each component of the AMS-CycleGAN model plays a crucial role in advancing the image quality to a higher level.

### D. COMPREHENSIVE ANALYSIS

In this section, we demonstrate the generalization capability of the AMS-CycleGAN model by conducting training on the summer2winter, orange2apple, and photo2flower datasets, as illustrated in Fig.12. During the image transformation process, the network learns the correlations between the two domains as well as the distinctive characteristics of each domain. These results indicate that the generated images by the AMS-CycleGAN model preserve the structural information of the input images while also capturing the shapes and characteristics of the target domain. For instance, in Fig.12 (a), the transformation from summer to winter results in the green trees in the first row being converted into bare trees covered with ice and snow. Additionally, in Fig.12 (b), the oranges are transformed into apples while the background remains unchanged. In the third row of Fig.12 (b), the presence of leaves partially occludes the oranges, yet the generated apples align with the expected output. Lastly, the ability of the generated model to achieve image focusing is illustrated in the third row of Fig.12 (c), as the flowers remain sharply focused despite the blurred background.

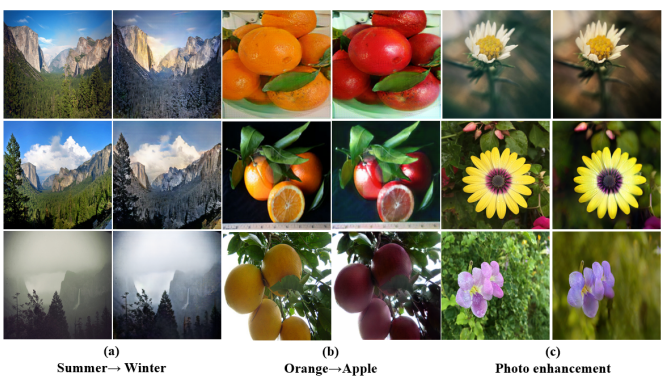


**FIGURE 12.** Generalization experiments of our proposed method on different datasets.

## VII. CONCLUSION

In this paper, a novel network called AMS-CycleGAN was proposed for style transfer. Firstly, the network incorporates the PONO-MS module between the decoder and encoder of the generator to preserve the structural information from the input. Secondly, the MS-SSIM loss is introduced in the reconstruction loss to strengthen the constraints on image brightness, colour contrast, and structural aspects in the generated images. Lastly, a channel-wise attention mechanism is added to the discriminator to guide the generator in emphasizing the crucial content of the input image. The effectiveness of the AMS-CycleGAN network is further confirmed through qualitative and quantitative experiments, demonstrating that the generated images exhibit enhanced perceptual visual quality and more comprehensive semantic information. In conclusion, the improvements made to the generator, discriminator, and loss functions bring meaningful advancements to style transfer. In the future, we will focus on generating high-quality images by lightweight networks.

## REFERENCES

[1] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 11, pp. 3365–3385, 2019.

[2] L. Gatys, A. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Twenty-Ninth Annual Conference on Neural Information Processing Systems*, 2016, pp. 262–270.

[3] P. Wilmot, E. Risser, and C. Barnes, "Stable and controllable neural texture synthesis and style transfer using histogram losses," *ArXiv*, vol. abs/1701.08893, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:10866737

[4] S. Li, X. Xu, L. Nie, and T.-S. Chua, "Laplacian-steered neural style transfer," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1716–1724.

[5] C. Castillo, S. De, X. Han, B. Singh, A. K. Yadav, and T. Goldstein, "Son of zorn's lemma: Targeted style transfer using instance-aware semantic segmentation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 1348–1352.

[6] C. Li and M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2479–2486.

[7] A. J. Champandard, "Semantic style transfer and turning two-bit doodles into fine artworks," *CoRR*, vol. abs/1603.01768, 2016. [Online]. Available: http://arxiv.org/abs/1603.01768

[8] Y.-L. Chen and C.-T. Hsu, "Towards deep style transfer: A content-aware perspective," in *British Machine Vision Conference*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:30329574

[9] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang, "Visual attribute transfer through deep image analogy," *ACM Transactions on Graphics (TOG)*, vol. 36, pp. 1 – 15, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:7685985

[10] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[11] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture networks: feed-forward synthesis of textures and stylized images," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, 2016, pp. 1349–1357.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*, vol. abs/1511.06434, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:11758569

[13] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *arXiv preprint arXiv:1610.07629*, 2016.

[14] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stylebank: An explicit representation for neural image style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1897–1906.

[15] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[18] J. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," *CoRR*, vol. abs/1711.11586, 2017. [Online]. Available: http://arxiv.org/abs/1711.11586

[19] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.

[20] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 551–10 560.

[21] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in neural information processing systems*, vol. 30, 2017.

[22] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *CoRR*, vol. abs/1607.08022, 2016. [Online]. Available: http://arxiv.org/abs/1607.08022

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[24] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5880–5888.

[25] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[26] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[27] J. Kim, M. Kim, H.-W. Kang, and K. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," 07 2019.

[28] L. Wen, C. Gao, and C. Zou, "Cap-vstnet: Content affinity preserved versatile style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 300–18 309.

[29] R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang, "Reusing discriminators for encoding: Towards unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8168–8177.

[30] B. Li, F. Wu, K. Q. Weinberger, and S. Belongie, "Positional normalization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[31] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[32] H. Taud and J. Mas, "Multilayer perceptron (mlp)," *Geomatic approaches for modeling land change scenarios*, pp. 451–455, 2018.

[33] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[34] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[35] S. Wu, G. Li, L. Deng, L. Liu, D. Wu, Y. Xie, and L. Shi, "$l1$-norm batch normalization for efficient training of deep neural networks," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 7, pp. 2043–2051, 2018.

[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[37] T. Salimans, H. Zhang, A. Radford, and D. Metaxas, "Improving GANs using optimal transport," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rkQkBnJAb

[38] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

XIAODI RANG was born in 1997. She received the bachelor's degree in software engineering from School of Computer and Control Engineering, Yantai University, in 2020. She is currently pursuing the master's degree with the School of Computer and Control Engineering, Yantai University. Her main research interest includes image processing and deep learning.

ZHENGYU DONG was born in 2000. He receive the bachelor's degree in software engineering from the School of Computer and Control Engineering at Yantai University in 2022. He is currently pursuing a master's degree in Computer and Control Engineering at Yantai University. The current main research direction is image style transfer.

JIACHEN HAN was born in 1998. he received the bachelor's degree in software engineering from School of Information Technology, Luoyang Normal University, in 2021. he is currently pursuing the master's degree with the School of Computer and Control Engineering, Yantai University. His main research interest includes image processing and deep learning.

CHAOQING MA received the B.S. degree in Computer Science and Engineering from Ludong University, China, in 2010, the M.S. degree and Ph.D. degree in Computer Science and Engineering from Chonbuk National University, South Korea, in 2012 and 2016, respectively. She is currently a lecturer in School of Computer and Control Engineering at Yantai University, China. Her research interests are in the areas of biological modelling and simulation, medical image processing and patter recognition etc.

GUANGZHI ZHAO received the M.S degree in Computer Science and Engineering from Jeonbuk National University, South Korea, in 2016, respectively. He is currently pursuing Ph.D in Computer Science and Engineering from Jeonbuk National University, South Korea. His research interests are in the areas of crop diseases and pests image recognition and medical image recognition etc.

WENCHAO ZHANG was born in 1993. He received his Master's degree in computer technology from the School of Computer Science and Engineering of Northwest Normal University in 2015. His main research interests include algorithm design and analysis.