Adaptive Content Feature Enhancement GAN for Multimodal Selfie to Anime Translation

Yuanming Li[1]
lym7499500@korea.ac.kr
Youngsaeng Jin[1]
youngsjin@korea.ac.kr
Jeonggi Kwak[1]
kjk8557@korea.ac.kr
Dongsik Yoon[1]
kevinds1106@korea.ac.kr
David Han[2]
dkh42@drexel.edu
Hanseok Ko*[1]
hsko@korea.ac.kr

[1] Korea University
Seoul, South Korea
[2] Drexel University
Philadelphia, USA

Abstract

With astonishing successes of GAN-based style transfer techniques, real-world photo translation to animation style images recently has attracted some interest. In particular, the transfer of a selfie to a cartoon has become quite popular as it can serve as a cartoon filter in social media. Unlike most of the Image-to-Image translation tasks, the selfie to anime task requires preserving the contour in the selfie image in the transfer process while it transforms other local characteristics into an animation style. Since the gap between the selfie domain and anime domain is quite large, as it can be imagined in the case of transforming a person into a cartoon animal like a mouse, developing an effective method remains a difficult challenge. In this paper, we propose an Adaptive Content Feature Enhancement Generative Adversarial Networks (ACFE-GAN) for a selfie to anime translation. By our model, the preservation of content features of selfie is improved. In addition to facial and hair contours, the shapes of worn items (*e.g.* hat and glasses) are also better preserved compared to existing methods. Our method also captures local features more accurately and selectively in translating them into the animation domain. Moreover, compared to the previous photo to anime translation models, we implement it with multimodal translation. Experiments on the selfie2anime dataset demonstrate that our method delivers superior performance in terms of selective preservation of content features.

Figure 1: Selfie to anime translation results by the proposed model. The first row shows the reference anime-face. The second row shows the input image and its translated images.

1 Introduction

Image-to-Image (I2I) translation is defined as the problem of converting the representation of one image into another. This problem is related to style transfer, which transforms the style of a source image to that of the target domain while preserving the content of the source image. Transforming photos of real-world scenes (or selfies) into animation style is a challenging task in computer vision as it includes the element of artistic style encompassing large diversities. Hence, there has been a large gap between these two domains.

Generative adversarial networks (GAN) [8] based image translation have achieved impressive results. Pix2Pix [13] is a type of conditional GAN, or cGAN, where the generation of the output image is conditional on an input image. Recently, numerous conditional GAN-based methods have been proposed, which implement I2I translation among the various domains, such as colorization [22, 28], image inpainting [27, 30] and photo to anime[4, 15], as well as to other domains like videos and 3D data. Compared to other I2I translation tasks, selfie to anime translation not only need to change the rendering but also needs to make significant variations of local characteristics (*e.g.* eyes or mouth). Existing algorithms employed perceptual loss [14] or cycle consistency loss to preserve the content feature. However, it is difficult to selectively convert some local characteristics into specific forms of animation styles while preserving the global characteristics of the face.

Previous methods [2, 4] utilized perceptual loss to preserve content features of the input image by minimizing the feature space distance between the two images. This constraint limits the method's effectiveness in cases of exaggerated animation style (*e.g.* big eyes). Therefore, in the U-GAT-IT [15], the perceptual loss is replaced by adaptive layer instance normalization (AdaLIN). The AdaLIN allows the transformation of some parts of the content features, such as shape or texture. However, U-GAT-IT based methods lack diversity in their outputs due to one-to-one mapping.

In this paper, we introduce a novel adaptive content feature enhancement (ACFE) block applied the selfie to anime translation framework to address the problem mentioned above. Figure 1 shows some examples of the multimodal selfie-to-anime translation generated by our proposed method. The input image is translated into different animation styles with reference-guided synthesis. In order to flexibly enhance the content information of an input image, we extract the content information from the encoder at different stages of the encoding process using skip connections. At each stage, the proposed ACFE block selectively abstracts content features and preserves them in the decoding process by scaling and

shifting processes. Based on our literature survey, our proposed ACFE structure is the first of its kind in adaptively abstracting content information from the encoders. It's been shown that our proposed ACFE-based method generalizes well even when the input image is from a different domain compared to the training set. The effectiveness of the proposed model is validated by extensive evaluations with different state-of-the-art multimodal I2I translation models [1, 5, 12, 19].

Our main contributions can be summarized as follows:

- We propose a GAN-based framework, note as ACFE-GAN, that maps selfies into the anime domain by an unsupervised multimodal translation.
- We developed a novel adaptive content feature enhanced block to selectively preserve the content information of input images.
- Extensive experiment results show that the proposed method can translate selfies to high-quality and diverse anime-faces in comparison to state-of-the-art multimodal translation models.
- The results of extra dataset with pre-trained models validate the generalizability of our approach.

2 Related Work

2.1 Style Transfer

In the recent past, inspired by convolutional neural networks (CNNs), Gatys *et al*. [7] proposed a CNN based style transfer model for turning a photo into a painting. An iterative optimization method was used to achieve the objective of matching desired feature distributions, which contains both the content feature of the source image and the style feature of a target image. Another similar method for style transfer is proposed by Huang *et al*. [11]. In their approach, the conditional instance normalization is modified to an adaptive instance normalization (AdaIN). The main idea of AdaIN is to align the mean and variance of the content image features to the mean and variance of the style image features. However, the method has been shown to be limited in generalizing to unseen styles.

More recently, Li *et al*. [20] attempted to transfer arbitrary style by using a series of feature transformations. A pre-trained VGG is employed as an encoder and the training is done only on the corresponding decoder. Instead of using the AdaIN layer, a pair of whitening and coloring transformations (WCT) is used. Their method is based on the observation that the whitening transformation can remove style-related information (*e.g*. color, texture) and at the same time retain the content features. Then comes the coloring operation, which is the inverse of the whitening operation. This is equivalent to integrating the style map into the generated image when restoring the original image.

2.2 Image-to-Image Translation

Supervised I2I translation aims to translate input images into the target domain with paired images as the source domain and target domain for training. Isola *et al*. [31] first proposed a conditional GAN-based model which is named pix2pix to solve a unimodal supervised I2I task. The pixel-wise L1 loss is an extra constraint between the output image and the ground truth. BicycleGAN [32] is the first supervised multimodal I2I model which is trained to translate input images into a target domain output with diverse styles. The network is

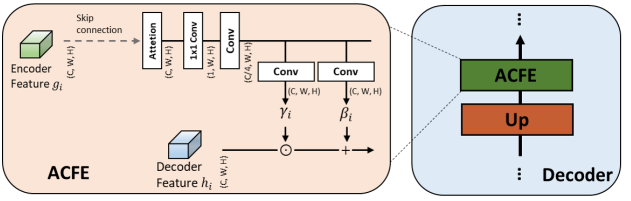


Figure 2: In the adaptive content feature enhancement (ACFE) block, the encoder feature g_i is first projected onto an embedding space with only one channel. And then increase the channel of feature by convolution. Finally, convolved to produce the modulation parameters (γ_i, β_i). These parameters are element-wise multiplied and added to the decoder feature h_i.

composed of cVAE-GAN [18] and cLR-GAN [3], to learn a set of solutions to mitigate the mode collapse problem.

Unsupervised I2I translation uses two unpaired sets of training images to convert images from one domain to another domain. CycleGAN [31], DiscoGAN [16] and DualGAN [26] exploit a novel constraint of the cycle consistency loss in the unsupervised GAN framework. The key idea of cycle consistency loss is that consistency is maintained between the source image and its reconstruction image. UNIT [21] assumes that two domains map to a shared latent space and learns a unimodal mapping. Moreover, MUNIT [12] and DRIT [19] refine the latent code into content code and style code on the basis of UNIT. Images of different domains share content space while their style space may not overlap significantly. Combining content coding with different style coding can get more robust and diverse results.

3 Proposed Method

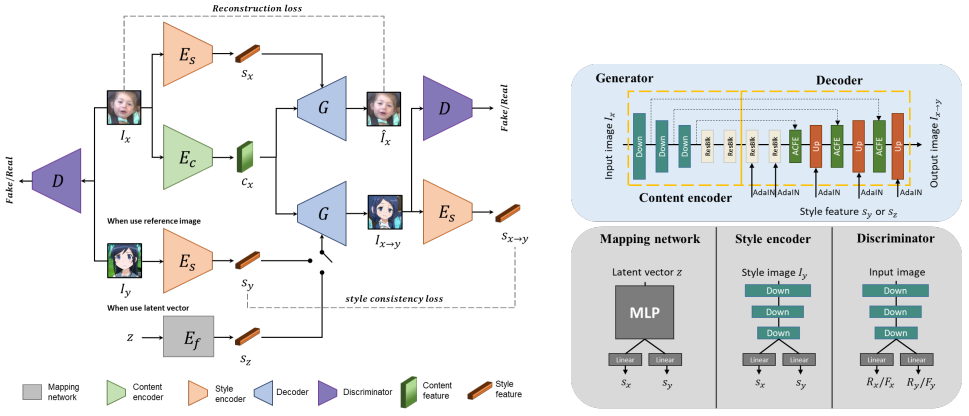
3.1 Adaptive Content Feature Enhancement Block

As mentioned in the introduction, perceptual loss or cycle consistency loss may put too much constraint on style feature transfer in anime domain. Hence, we design an adaptive content feature enhancement block, which is inspired by spatial feature transform [24], to enhance the content feature adaptively without degradation of the style information.

As show in Figure 2, the ACFE block is used to learn a mapping function Q that outputs a modulation parameter pair (γ_i, β_i) based on the features g_i from the content encoder $(\gamma_i, \beta_i) = Q(g_i)$. After obtaining (γ_i, β_i), we scale and shift decoder features by these modulation parameters:

$$ACFE(h_i, \gamma_i, \beta_i) = \gamma_i h_i + \beta_i \tag{1}$$

where h_i denotes the decoder feature tensor of the i-th layer, they are fed to the upsampling process. We aggregate content information of encoder feature by using an attention module similar to convolutional block attention module (CBAM) [25]. In CBAM, the encoder feature passes through the channel attention [10] and spatial attention in sequence. In the channel attention part, we discard the max-pooling operations to focus on the global structure. After the attention block, we apply the 1×1 convolutional layer to reduce the number of channels to 1, so that it can extract the feature as it only contains the content related information.



(a) (b)

Figure 3: (a) Overview of proposed model. (b) Architecture of content encoder, style encoder, mapping network, decoder and discriminator.

3.2 Model Overview

Figure 3(a) depicts the basic framework of our method. Let $I \in I_x, I_y$ represent a sample from the source domain X and the target domain Y. The overall architecture has the following structure: the content encoder E_c extracts the content features c_x from the input image I_x and feed it to decoder G. AdaIN [6, 11] injects style features s_y or s_z into G to generate images $I_{x\to y}$. The discriminator D will then determine whether the image is from the real samples or synthetic samples. We only use one discriminator to distinguish true and false images from two domains. The style features s_z and s_y are provided either by the mapping network E_f or by the style encoder E_s.

The main part of our model is an auto-encoder architecture which is composed of encoders and a decoder as shown in Figure 3(b). The content encoder and style encoder extract the content features and style features, respectively. The decoder is then able to map an image by combing content and style features. The ACFE blocks are inserted after the upsampling layers in the decoder part to enhance the content feature of the source image. Style encoder, mapping network, and discriminator constitute the shared network part and domain-specific mapping part. The domain-invariant features of both domains are extracted by the shared network part. Then, remapping the features to the corresponding domain is done by domain-specific part with two branches of fully connected layers. The shared network part can encourage the generalization to unseen samples [5].

3.3 Loss Function

The full objective of our model comprises four loss functions as follows:

Adversarial Loss. An adversarial loss is employed to match the distribution of the translated images to the target image distribution as shown by the following equation.

$$L_{adv} = E_Y[\log(D_Y(I_y))] + E_{X,Y}[\log(1 - D_Y(\hat{I}_{x\to y})] \tag{2}$$

where $D_Y(.)$ is the output of discriminator corresponding to the target domain Y. Generating output image $\hat{I}_{x\to y}$ with content feature c_x can be done in two different ways: (1) style feature

s_y from style encoder E_s or (2) style feature s_z from mapping network E_f.

Reconstruction Loss. The reconstruction loss is defined as the L_1-norm of the distance between original and reconstructed images.

$$L_{rec} = E_X[\|I_x - \hat{I}_x\|_1] \tag{3}$$

where $\hat{I}_x$ is generated by combination of content feature c_x and style feature s_x of input image I_x. The image reconstruction loss is used to make sure the generator can reconstruct the original image within the domain.

Style Consistency Loss. The style consistency loss is applied to ensure that the style encoder encodes meaningful style features s_y for output image $\hat{I}_{x\to y}$. The loss function is defined as follows.

$$L_{sty} = E_Y[\|s_y - \hat{s}_y\|_1] \tag{4}$$

where $\hat{s}_y = E_s(\hat{I}_{x\to y})$. When using the style feature s_z ob[tined from mappin] network E_f. The style consistency loss is change by $E_z[\|s_z - \hat{s}_z\|_1]$, where $\hat{s}_z = E_f(\hat{I}_{x\to y})$.

Mode Seeking Loss. The mode seeking loss is employed to enhance the output diversity as follows:

$$L_{ms} = -E_{X,Y}[\|\hat{I}^1_{x\to y} - \hat{I}^2_{x\to y}\|_1] \tag{5}$$

where the $\hat{I}^1_{x\to y}$ and $\hat{I}^2_{x\to y}$ are translated images with different style feature. When the generator cannot produce diverse outputs, the distance between them will vary small. Thus, maximizing this term can increase the diversity of output.

Full Objective. Our full objective functions can be summarized as follows:

$$\begin{aligned} L_D &= L_{adv} \\ L_G &= L_{adv} + \lambda_{rec}L_{rec} + \lambda_{sty}L_{sty} + \lambda_{ms}L_{ms} \end{aligned} \tag{6}$$

where λ_{rec}, λ_{sty} and λ_{ms} are hyperparameters for each term. The all these hyperparemeters are set to 1.

Optimization. We use Adam optimizer [17] with parameters $\beta_1 = 0$ and $\beta_2 = 0.99$, and the batch size is set to 7. The initial learning rate of generator and discriminator are both 0.0001. We obtain the best performance before 100,000 iterations on a NVDIA 3090 GPU.

4 Experiments

4.1 Baseline

We have compared our method with various state-of-the-art multimodal image translation models including MUNIT[12], DRIT[19], DSMAP[1] and StarGAN v2[5]. We used the official code of all baseline methods. All models including our model are bidirectional translation models. Since this study only focuses on selfie to photo mapping, the results of anime to selfie mapping are not included.

Figure 4: Samples of selfie2anime dataset.

4.2 Dataset

We evaluate our model on the selfie2anime dataset which contains 3,400/100 images on each domain for training/testing.The dataset is publicly available in Kaggle and unpaired. As shown in Figure 4, the selfies have a variety of poses and views, and the backgrounds are complex. In contrast, anime face domain are more standardized in these aspects. The disadvantage of anime face domain is that it often lacks some recurring items in selfies (*e.g.* phone, hat). As such, the task here is more challenging. Moreover, we use some extra data from the CelebA-HQ dataset as test images to validate the generalization of our model. All images are resized to 128×128 resolution for training and testing.

4.3 Evaluation

We generated 1,000 images using all images from the test dataset as input images. We use each image to produce 10 generated images with different styles. We conduct quantitative evaluations using the following metrics.

FID. We use FID [9] to measure the distance distributions between the generated images and real images through features extracted by Inception Network [23]. Lower values of FID indicate the distributions between generated images and real images are closer and better quality of the generated images is produced.

LPIPS. We employ LPIPS [29] to evaluate the diversity of generated images. LIPIS measures the average feature distances among generated images. Higher values of LPIPS indicate the generated images are more diverse.

MOS. We use mean opinion score (MOS) as a subjective evaluation metric. We asked 50 raters to assign an integral score from 1 (bad quality) to 5 (excellent quality) to the translated images. Each rater votes for 30 images (6 sets of translated images). Each set of images is produced by the same selfie.

5 Results

Comparison on Content Preservation. Figure 5 illustrates a qualitative comparison of the baseline models and our model. Each model translates the selfie to anime with random style features which is the output of the mapping network. We observe that generated images of our model has a higher visual quality than generated images of other models. For the global features (*e.g.* contour of face and hair), MUNIT, StarGAN v2, and our model have better representation. For the some local features (*e.g.* glasses and hat), only our model successfully preserves contour of these items. Although MUNIT and StarGAN v2 can generate relatively

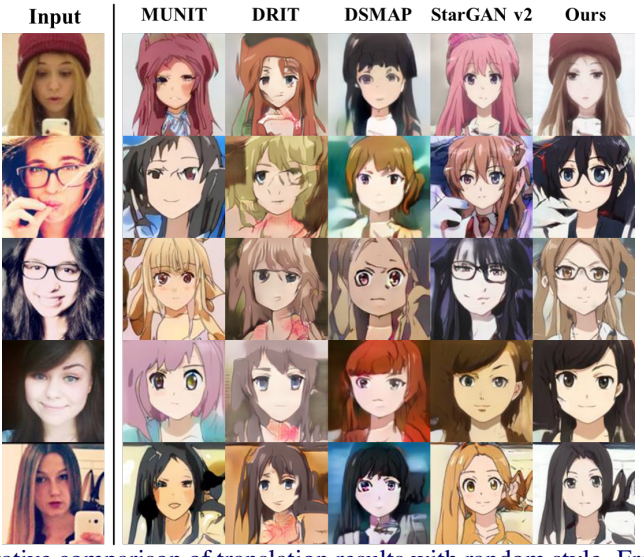


Figure 5: Qualitative comparison of translation results with random style. From left to right: input selfie, MUNIT, DRIT, DSMAP, StarGAN v2 and our model.

Method	MUNIT	DRIT	DSMAP	StarGAN v2	**Ours**
FID	102.1	123.4	106.4	101.7	**96.8**
LPIPS	0.4052	**0.2287**	0.3670	0.3624	0.3245

Table 1: FID and LPIPS results compare to other baselines. A low FID indicates high visual quality. A low LPIPS indicates that the translated images will not have structural changes depending on the different reference images.

well global features. For some local features, it is difficult for them to preserve the same performance as global features. In addition, MUNIT, DRIT and DSMAP generate diversity visible artifacts in output images.

As shown in Table 1, our model achieves the lowest FID compared to the baselines. It indicates the synthetic samples generated by our model are more similar to the real dataset than the baselines. Although DRIT obtains the lowest LPIPS, the quality of generated images is undesirable. We hope that the content information remains unchanged as much as possible, while the style information will change with different Styles (*e.g.* color and eyes). Compared to the change of style feature, the change of content feature have a greater impact on the value of LPIPS. Thus, a low LPIPS is benefit for us. The result of LPIPS of our model is not the lowest, but this does not mean that our model has a poor performance on preservation of the content feature. As shown in Figure 6, DRIT suffer from mode-collapse, which results in lower LPIPS. The second-lowest of LPIPS is our approach. This shows that our model is effective in the preservation of content features. With different reference images, our model still preserve well content features of input images.

Comparison on Reference-Guided Synthesis. As shown in Figure 6, we use three specific images as reference to generate corresponding images. It is obvious that generated images of our model have the best visual quality compared to the other baselines. The baseline methods hardly reflect the content information of the input image and they just match the

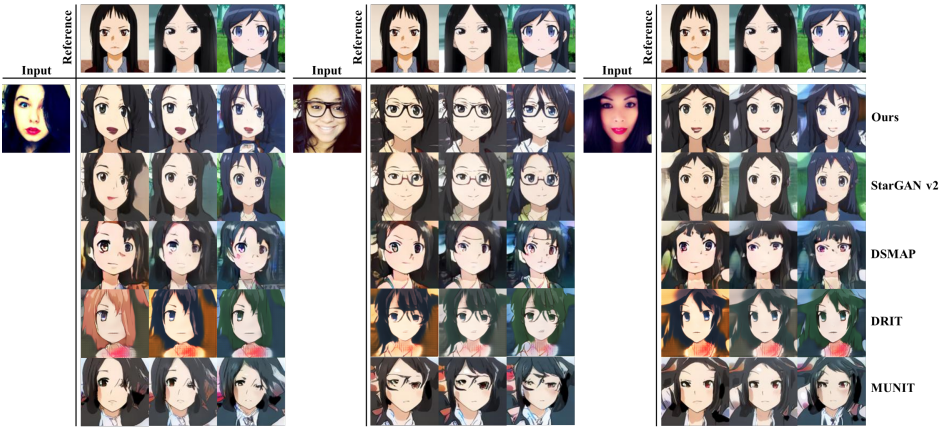
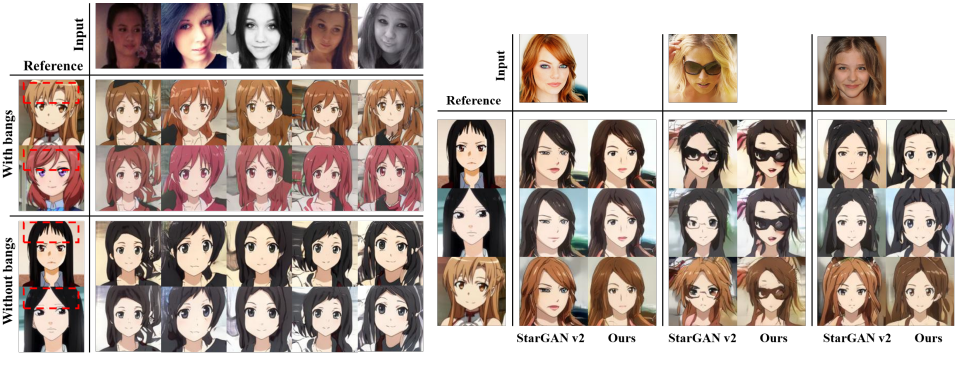


Figure 6: Qualitative comparison of reference-guided translation results. Each model translates the input selfie into anime domain and reflecting the styles of the references images.



(a) (b)

Figure 7: (a) Visual results of our model using reference images w/ and w/o bangs. (b) Qualitative comparison of reference-guided translation results on CelebA-HQ dataset with the model trained on selfie2anime dataset.

input to the anime domain. In contrast, our model not only preserve content feature but also correctly match color from reference images. Figure 7(a) further demonstrates our model can map selfie images into animation style images that reflect the various styles of the reference images. In addition, the color is reflected in the output images, the bangs of reference image also effect the translated results. If there are bangs in the reference image, the model will also output an image with bangs.

User Study. We conducted mean opinion score (MOS) evaluation on the pictures output by each model. A mean opinion score (MOS) result is shown in Table 2, indicating our method significantly outperforms baselines, both in terms of image quality and preservation of content of input image. We confirmed the superior perceptual performance of our method using MOS testing.

Generalization on Other Dataset. In order to demonstrate the generalizability of our model, we use a pre-trained model which is trained on the selfie2anime dataset to test on CelebA-HQ dataset. As shown in Figure 7(b), StarGAN v2 and our model translate the

Method	MUNIT	DRIT	DSMAP	StarGAN v2	**Ours**
MOS	1.89	2.19	1.92	3.10	**3.55**

Table 2: MOS score for each model. For each method 1500 samples (30 images × 50 raters) were assessed.

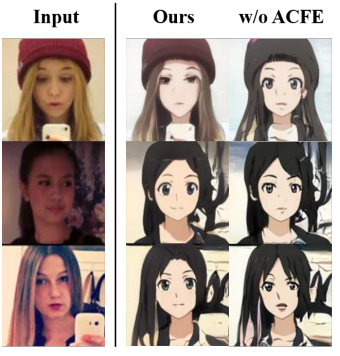


Figure 8: Qualitative comparison of ablation study.

Method	Proposed	w/o ACFE
FID	**96.8**	98.9
LPIPS	**0.3245**	0.3439

Table 3: FID and LPIPS results of ablation study.

photo-face into anime-face with reference-guided synthesis. In spite of the distribution gap between the two datasets, our model successfully generates superior images and maintains high similarity of content feature of the input images. In some details of generated images, our model improves the preservation of content information. However, there is a drawback with color-shifting between the reference images and generated images. These results show that our model has generalizability on unseen samples.

Ablation Study. The proposed ACFE block clearly transfers some key content features correctly compared to the result without it. As shown in Figure 8, the first figure transfers the fold of the beanie cap from the input correctly while the image without ACEF block doesn't. The flat forehead shape of the input on the second and the third row images were well reflected on the images created with ACEF block while the ones without failed to do so. Apart from this, the results of FID and LPIPS show that the image quality generated by our model is more superior as shown in Table 3.

6 Conclusions

In this paper, we presented a GAN-based framework for multimodal selfie to anime translation. We proposed a novel ACFE block on the decoder to adaptively preserve content features of source images. The experimental results showed that the proposed method handles selfie to anime translation better than the state-of-art methods in terms of accurately preserving content features. Moreover, our model exhibited an generalization as demonstrated by its experimental results on another datasets of different domain.

Acknowledgment

This research was supported by Deep Machine Lab (Q2109331).

References

[1] Hsin-Yu Chang, Zhixiang Wang, and Yung-Yu Chuang. Domain-specific mappings for generative adversarial style transfer. In *ECCV*, pages 573–589. Springer, 2020.

[2] Jie Chen, Gang Liu, and Xin Chen. Animegan: A novel lightweight gan for photo animation. In *International Symposium on Intelligence Computation and Applications*, pages 242–256. Springer, 2019.

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. 2016.

[4] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. CartoonGAN: Generative adversarial networks for photo cartoonization. In *CVPR*, pages 9465–9474, 2018.

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[6] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 3(7):e11, 2018.

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016.

[8] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. 2017.

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017.

[12] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[15] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *ICLR*, 2020.

[16] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865. PMLR, 2017.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, pages 35–51, 2018.

[20] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *arXiv preprint arXiv:1705.08086*, 2017.

[21] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. 2017.

[22] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *CVPR*, pages 7968–7977, 2020.

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[24] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, pages 606–615, 2018.

[25] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, pages 3–19, 2018.

[26] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2849–2857, 2017.

[27] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, pages 5505–5514, 2018.

[28] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, pages 649–666. Springer, 2016.

[29] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[30] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021.

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.

[32] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, pages 465–476. 2017.