2023 IEEE International Conference on Big Data and Smart Computing (BigComp) | 978-1-6654-7578-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/BigComp57234.2023.00041

Enhancing Image Representation in Conditional Image Synthesis

Jonghwa Shim
School of Electrical Engineering
Korea University
Seoul, Republic of Korea
indexlibrorum3822@korea.ac.kr

Eunbeen Kim
School of Electrical Engineering
Korea University
Seoul, Republic of Korea
gichanac@korea.ac.kr

Hyeonwoo Kim
School of Electrical Engineering
Korea University
Seoul, Republic of Korea
guihon12@korea.ac.kr

Eenjun Hwang
School of Electrical Engineering
Korea University
Seoul, Republic of Korea
ehwang04@korea.ac.kr

***Abstract*—Even though deep neural network-based conditional image synthesis has shown impressive advances in terms of image quality, they still fall short of dealing with domain-dependent global and local styles and distinct shape representations of synthesized images. To address this issue, we propose a novel GAN-based conditional image synthesis model that incorporates a conditional normalization layer called IAN for style and edge-weighted shape enhancing loss for shape. Comparative experiments and ablation studies on popular and different domain datasets show that the proposed model outperformed other popular image-to-image translation model for diverse image domains.**

***Keywords*—Generative Model, Conditional Image Synthesis, Image Representation, Normalization Layer, Edge Detection**

I. Introduction

Conditional image synthesis refers to the generation of photorealistic images for condition images given as input. Traditional methods for this have been to concatenate image fragments or use image collection to compute the resulting image [1]. Recently, deep neural networks have been used to directly learn mappings between inputs and outputs [2,3]. In particular, Generative Adversarial Networks (GANs)-based methods have attracted much attention due to their versatility and superior quality of generated images. Depending on the domain of the input data and the output image, conditional image synthesis can be defined by various tasks [4,5] such as colorizing, transforming styles, and editing images.

To generate good quality of images for conditional images (i.e., image-to-image translation), we need to consider two types of semantic elements in images: style and shape. Style representation refers to spatially and semantically appropriate color expression and surface pattern. On the other hand, a shape representation describes the structural shape of an object in an image. Recent studies on style representation have used various conditional normalization layers that work for style transfer and high-resolution image synthesis [6,7,8]. These conditional normalization layers effectively propagate semantic information of input data to neural network by adaptively modulating intermediate feature maps. On the contrary, works on shape representation have used edge maps to increase the clearness of output images in super-resolution tasks [9,10] or utilize perceptual loss function [11] based on a pretrained network to preserve image content. Despite a lot of efforts so far, it is still challenging to properly handle both style and shape representations in conditional image synthesis. Specifically, in style representation, it is difficult to effectively control both the global and local styles of an image. Likewise, in shape representation, edge maps are difficult to use except for super-resolution tasks, and the perceptual loss function cannot guarantee performance in domains where the pretrained network has not learned [11].

In this paper, we propose a novel GAN-based conditional image synthesis model based on improved style and shape representations. For style representation, we propose a conditional normalization layer called integrated adaptive normalization (IAN). IAN enhances both global and local style representations by combining functional strengths of existing conditional normalization layers. For shape representation, we propose an edge-weighted shape enhancing loss to make the shape of generated images clearer by using an edge map of real images extracted by canny edge detector. This loss function has advantages in terms of model complexity and ease of application as it does not require a separate pretrained network for edge information. As these two methods have minimal impact on each other's purpose (style and shape representation), both can be easily applied to model training simultaneously. Various comparative experiments that we carried out using popular datasets such as Cityscapes, CelebAMask-HQ, and High-resolution anime show that our proposed model outperforms other existing techniques. In addition, we demonstrate the effectiveness of the proposed model through extensive ablation studies.

The main contributions of this paper are as follows.

- We propose a conditional normalization layer called IAN to handle global and local texture representations effectively.
- We show how to use edge-weighted shape enhancing loss to improve shape clarity in generated images.
- We verify the performance and impact of the proposed method through comparative experiments and ablation study using various datasets.

The structure of this paper is as follows. In Section 2, we introduce several related works. Section 3 describes in detail the proposed model and Section 4 verifies the performance and impact of the proposed model through comparative experiments and ablation study. Finally, Section 5 concludes this work.

This research was financially supported by the Ministry of Trade, Industry and Energy(MOTIE) and Korea Institute for Advancement of Technology(KIAT) through the International Cooperative R&D program. (Project No. P0017192)

978-1-6654-7578-5/23/$31.00 ©2023 IEEE
DOI 10.1109/BigComp57234.2023.00041

II. Related Works

2.1 Deep Generative Models

Recently, various deep generative models such as generative adversarial networks (GANs) [12] and variational autoencoders (VAE) [13] have shown good performance in generating realistic images compared to other approaches. GAN consists of a generator and a discriminator. The generator tries to produce a realistic image so that the discriminator cannot distinguish the synthesized image from the real image. GANs generate random images by default, but conditional image synthesis is also possible by adding condition data of output images during training.

Depending on the type of input data, GANs for conditional image synthesis are divided into several types. For instance, the class conditional GANs [14] learns to synthesize images corresponding to the binarized category labels. GANs can also be fused with natural language processing models to generate images based on text [15,16]. Another interesting type of GAN is to perform image-to-image translation [17,18] using images as input and output.

2.2 Normalization Layers

The normalization layer was a key factor in the development of deep neural network structure. Normalization layers explicitly control activations, the intermediate outputs of the neural network, to improve performance. For instance, Batch Normalization (BatchNorm) [19] of Inception-v2 network stabilized the training of classification networks and improved their performance by preventing internal covariance shift. Since then, several normalization layers have been proposed for various purposes. For example, Instance Normalization (InstanceNorm) [20] bleaches the style information of the input image for style transfer. Layer Normalization (LayerNorm) [21] is structurally suitable for sequential models such as RNN. Group Normalization (GroupNorm) [22] is an alternative to BatchNorm to improve memory efficiency. These normalization layers are classified as unconditional normalization layers because they are not dynamic with respect to external condition data.

A conditional normalization layer, on the other hand, aims to inject information from external condition data into the network. Conditional normalization layers usually work in two steps. First, the output of each layer is normalized with zero mean and unit deviation. Second, the mean and standard deviation of the output are modulated using affine transformation inferred from the external condition data. Affine transformation inference uses additional neural network contained in the conditional normalization layer. For instance, Conditional Batch Normalization (Conditional BatchNorm) [8] and Adaptive Instance Normalization (AdaIN) [6] adjust the global style representation of the output. They were used for style transfer task and various vision work. Spatially Adaptive (De)Normalization (SPADE) [7] applied spatially-adaptive affine transformation using semantic maps. Similarly, Spatial Feature Transform (SFT) [23] proposed spatially dynamic feature modulation layers for super-resolution task. Overall, these works focused on the adjustment of either global or spatial(local) style representation. We combine the functional structure of existing conditional normalization layers to handle both styles.

2.3 Shape Enhancment

In image synthesis, it is important to properly reconstruct the clarity of the shape of an object. Popular reconstruction loss functions for image synthesis include mean-squared error (MSE), root mean-squared error (RMSE), and mean-absolute error (MAE). Although these pixel-wise mean error-based loss functions are generic, they do not correlate well with human perceived quality. This is because they do not consider salient features inherent in the image [9]. Therefore, simply optimizing these loss functions does not always lead to optimal perceptual shape representation of the output image. Recent approaches to address this problem include perceptual loss [11] and edge-based methods [9, 10]. Perceptual loss calculates the perceptual error between the output image and the target image based on the high-level features extracted from the pretrained VGG network [24] and replaces MSE loss in various computer vision tasks [25]. Edge-based methods enhance image clarity by using edge information, an intuitive alternative to shape representation. Several works on semantic segmentation added sub-branch networks to perform edge prediction for segmentation boundary refinement [26, 27]. Similarly, in the super-resolution task, edge detection branches were added to the GAN-based model to sharpen the high-frequency and shape information of the image [10, 28]. However, edge-based methods are limited to semantic segmentation or super-resolution tasks despite their rich shape representation of edges because the edges can optimize even very detailed shape information. So, we extract only prominent edges through edge thresholding and using them for image-to-image translation.

III. Methods

In this paper, we propose a method for improving the style representation and shape representation of generated images in GAN-based image-to-image translation. To do that, we first introduce a new conditional normalization layer, IAN to enhance style representation. Then, we describe our edge-weighted shape enhancing loss for shape representation. Lastly, we show the structure of our proposed model, loss functions, and additional elements for conditional image synthesis.

3.1 Integrated Adpative Normalization

IAN learns a mapping function that converts conditional images into realistic images, taking into account both global and local style representations. As in other normalization layers, activations are normalized and then modulated with the learned scale and bias. Let h^i be the activations of the i-th layer in a deep convolution network for a batch of N samples, and let C^i, H^i, and W^i be the number of channels, height, and width of h^i activations, respectively. Also, assume that $X \in L^{H\times W}$ represents a condition image with height H and width W. Then, the activations are normalized to unit distribution in channel-wise manner for $n \in N, c \in C^i, y \in H^i,$ and $x \in W^i$. The resulting normalized activations, o^i can be defined by Eq. (1)

$$o^i = \frac{h^i_{n,c,y,x} - \mu^i_c}{\sigma^i_c} \quad (1)$$

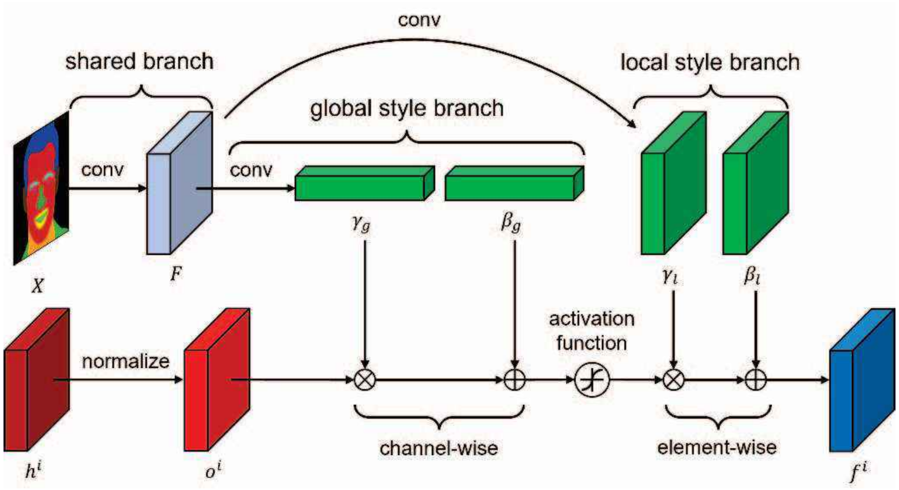

Fig. 1. Structure of IAN. IAN integrate the operations of existing normalization layers and effectively manipulates the global texture and local details simultaneously.

Here, μ_c^i and σ_c^i represent the mean and standard deviation of activations in channel c, respectively, and can be defined by the following equations.

$$\mu_c^i = \frac{1}{NH^iW^i} \sum_{n,y,x} h_{n,c,y,x}^i \quad (2)$$

$$\sigma_c^i = \sqrt{\frac{1}{NH^iW^i} \sum \left(\left(h_{n,c,y,x}^i \right)^2 - (\mu_c^i)^2 \right)} \quad (3)$$

As shown in Fig. 1, IAN consists of three branches: shared branch, global style branch, and local style branch. Each branch has three convolutional layers. The shared branch converts the condition image X into a shared style feature F. The shared style feature is passed as input to the global style branch and the local style branch to calculate the modulation parameters. The global style branch produces global style modulation parameters γ_g and β_g for scale and shift. Respectively. They are multiplied and added to the normalized activations in a channel-wise manner:

$$g^i = \gamma_g \circ \sigma^i + \beta_g \quad (4)$$

Here, $\circ$ indicates channel-wise multiplication and g^i denotes globally stylized activations. After adjusting the global style of the activations, the spatial and local style of the image need to be refined. As in the global style branch, the local style branch produces local style modulation parameters γ_l and β_l for scale and shift, respectively. They are multiplied and added to the globally stylized activations passed through the activations function ϕ in element-wise manner:

$$f^i = r_l \odot \phi(g^i) + b_l \quad (5)$$

Here, $\odot$ indicates element-wise multiplication and f^i denotes final stylized activations with global and local style representations for realistic image synthesis.

In fact, IAN integrates and generalizes several existing conditional normalization layers. For instance, if we remove the global style branch, reduce the number of convolution layers in the branch to one, and restrict the condition image to segmentation mask, then IAN becomes SPADE [7]. Similarly, if we remove the local style branch, replace the its convolutional layer with a fully-connected layer, and replace the condition image with a real image, IAN becomes ADAIN [6]. SPADE and ADAIN are complementary. In other words, the former is good at extracting rich spatial and local style representations from segmentation masks, while the latter is suitable for texture transformations that manipulate the global style of an image. The global and local branches of IAN accommodate both the advantages of conditional normalization layers using channel-wise or element-wise modulation. IAN also supports various condition image types (e.g., segmentation mask, sketch image). We used deeper convolutional layers (3 layers per branch) compared to SPADE for better performance [30]. In addition, we did the global stylization first and then the local stylization. That's because doing so has been more effective empirically.

3.2 Edge-weighted Shape Enhancing Loss

Images generated from image-to-image translations should retain their original intended shape. Although the adversarial loss function of GANs and the mean error-based reconstruction loss function described in Section 2.2 retain the shape information to some extent, they do not guarantee a satisfactory shape representation. Adversarial loss introduces undesirable distortion in the output image because it is difficult to converge to the optimal point [29] and mean error-based reconstruction loss does not consider key features of the image, resulting in blurriness in the image [9,12]. We use both shape information of real images and mean error-based reconstruction losses to guide the network to generate shape representations that are considered perceptually important. In particular, we use the edge map produced by the Canny edge detector as shape information.

Now, we describe our edge-weighted shape enhancing loss in detail. Basically, we focus on pixels belonging to the edge area in the mean error-based reconstruction loss. Let $Y \in L^{H \times W}$ and $\hat{Y} \in L^{H \times W}$ be the output image of the generator with height H and width W and the correct target image for the output image, respectively. In addition, $E \in L^{H \times W}$ is the binary edge map of the target image. As shown in Fig. 2, we can get the edge map E from the target image using the Canny edge detector [31]. In edge filtering, we set the low and high thresholds heuristically so that unnecessary edges are removed and only key edges are left. These threshold settings for edge extraction can vary by data domain. Then, the edge-weighted shape enhancing loss, $loss_{shape}$, is defined as the average of the element-wise multiplication of the binary edge map and the mean error-based reconstruction loss for $i \in H$ and $j \in W$:

$$loss_{shape} = \frac{\sum_{y=1}^{H}\sum_{x=1}^{W} E_{i,j} \odot (|\hat{Y}_{i,j} - Y_{i,j}|)}{HW} \tag{6}$$

Fig. 3 visualizes the procedure for calculating the shape enhancing loss. We chose MAE as the mean error-based reconstruction loss. The shape enhancing loss optimizes only the portion that corresponds to the shape representation of the image due to the edge map. The shape enhancing loss is then combined with the reconstruction loss, $loss_{recon}$, to give the pixel loss, $loss_{pixel}$ by Eq. 8.

$$loss_{recon} = \frac{\sum_{y=1}^{H}\sum_{x=1}^{W} (|\hat{Y}_{i,j} - Y_{i,j}|)}{HW} \tag{7}$$

$$loss_{pixel} = loss_{recon} + \lambda_{shape} loss_{shape} \tag{8}$$

Here, λ_{shape} adjust the ratio of the shape enhancing loss to reconstruction loss and we experimentally set λ_{shape} to 10. Pixel loss guides the network to produce sharper edges compared to conventional reconstruction loss. This method can be used for various image-to-image translation tasks with a minor modification and leads to more effective shape representation.

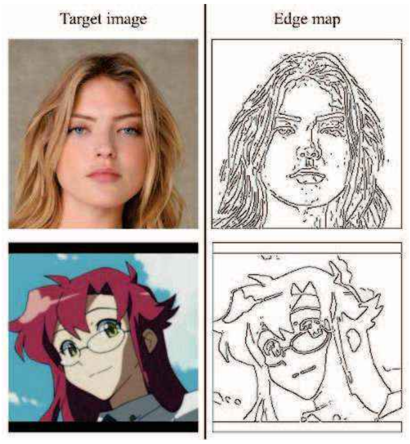

Fig. 2. Edge maps extracted from target images. The thresholds of first and second row are 5, 100 and 12, 100 respectively.

3.3 *Proposed Model*

In this section, we describe the overall structure of our proposed model. Firstly, we introduce GAN-based image-to-image translation network with IAN and then describes the loss functions including shape enhancing loss.

Fig. 4 shows the generator architecture of the proposed model. IAN encodes and feeds information about condition images at each layer. So, we can remove the encoder from generator of the encoder-decoder structure [32]. This simplification applies similarly to SPADE and ADAIN, which helps to make the model lighter [7]. Our generator consists of several ResNet blocks [33] and upsampling layers. As shown in the figure, IAN is implemented in the normalization layer of the ResNet block. As each ResNet block operates at a different scale, we downsample the input condition image to the IAN to match the resolution. The initial input of the generator is also the downsampled condition image.

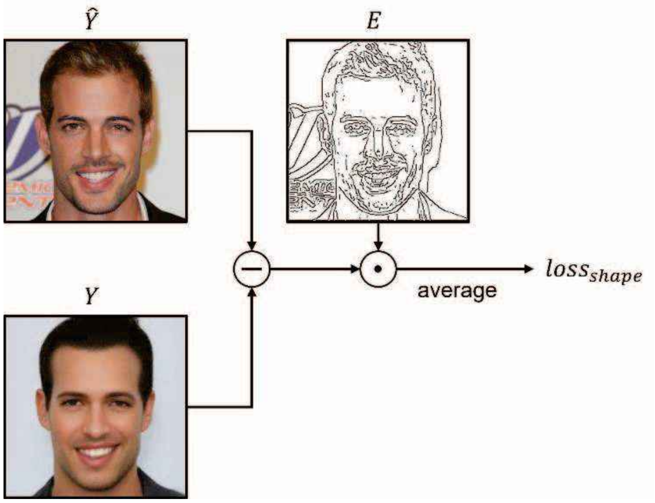

Fig. 3. Visualization of shape enhancing loss. In the edge map, the value of edge part is 1 and the rest is 0.

We also used the multi-scale discriminator of pix2pixHD [34], which consists of two discriminators: one receives the original input image and the other receives its half resolution image. Each discriminator simply consists of convolutional layers. We also applied spectral normalization [35] to all layers of the generator and discriminator. This modification contributes to stable model training.

Now, we describe the loss function that we used to train our network and its role. First, we consider the adversarial loss $loss_{adv}$ for training generator and discriminator. We chose the least square adversarial loss [36] with conditional input as the adversarial loss for generator G and discriminator D:

$$\min_{G}\max_{D} l\,oss_{adv} = E_{X\sim p_X(I),\hat{Y}\sim p_{\mathrm{dt}}(\hat{Y})}\left[\log D\left(\hat{Y}|X\right)\right] \\ +E_{I\sim p_I(I)}\left[\log\left(1 - D(G(X)|X)\right)\right] \tag{9}$$

In the proposed model, the least square loss experimentally performed better than the non-saturate loss [14] or Wasserstein loss [29]. In addition, we applied perceptual content loss, $loss_{content}$ for better shape representation [11].

$$loss_{content} = \frac{1}{C^i H^i W^i} |\delta^i(\hat{Y}) - \delta^i(Y)| \tag{10}$$

Here, δ^i is i-th intermediate layer of pretrained VGG network and C^i, H^i, W^i are the number of channels, height, and width of intermediate feature map $\delta^i(\quad)$, repectively. We employed VGG 19 [34] as our pretrained VGG network. Perceptual content loss optimizes shape information by considering the high-level features of the image. Although pixel loss alone achieved some sufficient performance, we found that perceptual content loss prevented some image synthesis failures. In addition we used feature matching loss $loss_{feat}$ [12], a popular training stabilization technique for GAN models:

$$\min_{G} loss_{feat} = |E_{x\sim p_{data}} D^i(\hat{Y}) - E_{z\sim p_z(z)} D^i(G(X))|_2^2 \tag{11}$$

Assuming D^i is i-th intermediate layer of the discriminator D, , the final loss function $loss_{final}$ is derived as:

$$loss_{final} = \lambda_{adv} loss_{adv} + \lambda_{pixel} loss_{pixel}$$

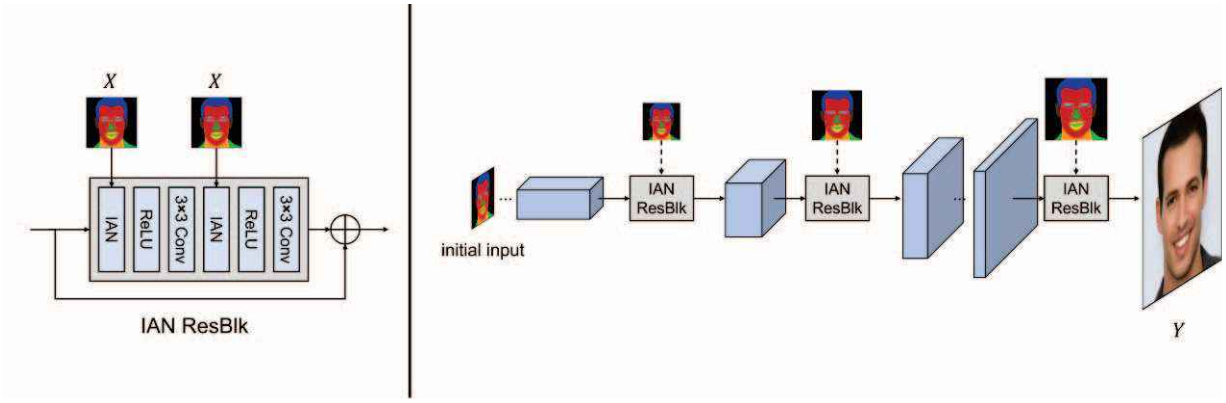

Fig. 4. (left) Structure of single residual block with IAN. IAN uses downsampled condition image X to modulate the activations. (right) Generator architecture of proposed model. Our generator consists of series connection of residual blocks and upsampling layers. The upsampling layer is omitted from the figure. Initial input of the first layer is also a downsampled condition image.

$$+\lambda_{content} loss_{content} + \lambda_{feat} loss_{feat} \quad (12)$$

Here, λ_q is the weight hyper-parameter for $loss_q$ and we set λ_{adv}, λ_{pixel}, $\lambda_{content}$, and λ_{feat} to 1, 1, 10, and 10, respectively.

IV. Experiment

4.1 Experimental setup

The learning rates of the generator and discriminator were set to 0.0001 and 0.0004 [36], respectively. We used ADAM [38] as the optimizer and set β_1 and β_2 to 0 and 0.999, respectively. All experiments were performed on NVIDIA RTX 3090 GPU. All models were selected at the highest performance point within the acceptable training period.

Datasets: We also used three different datasets to evaluate the performance of different image-to-image translation tasks. All images in the datasets were resized to 256 × 256 for the experiment. A brief description of the datasets used is as follows.

- Cityscapes dataset [39] is a large-scale database which focuses on semantic understanding of urban street scenes. It provides semantic and dense pixel annotations for 8 categories (flat surfaces, humans, vehicles, constructions, objects, nature, sky, and void). The dataset consists of around 3000 annotated images with the size of 256 x 256. Images were collected over several months in 50 cities during the day and in varying weather conditions. The low and high thresholds of the Canny edge detector for shape enhancing loss were set to 8 and 100, respectively. We used 2500 images for model training and 500 images for validation.
- CelebAMask-HQ[40] is a large-scale face image dataset that has 30,000 high-resolution face images selected from the CelebA-HQ[41]. Each image has segmentation mask of facial attributes corresponding to CelebA-HQ. The masks of CelebAMask-HQ were manually annotated with the size of 512 x 512 and 19 classes including all facial components and accessories. The low and high thresholds of the Canny edge detector for shape enhancing loss were set to 5 and 100, respectively. We used 29,700 images for model training and 300 images for validation.
- High-Resolution Anime Face Dataset [42] provides high-resolution Japan-style animation face images. This dataset was created by manually filtering low-quality images from the character anime illustration dataset Danbooru2019 Figures [42] and cropping only facial regions, including neckline, ears, and hats. All images are in 512 × 512 resolution. We extracted sketch images from animation images using line distiller and used the sketch images as condition images. The low and high thresholds of the Canny edge detector for shape enhancing loss were set to 12 and 100, respectively. We used 4,700 images for model training and 300 images for validation.

Performance metrics: The Fréchet inception distance (FID) [37] is used to assess the quality of images. Unlike the earlier inception score (IS) [43], which evaluates only the distribution of generated images, the FID compares the distribution of generated images with the distribution of real images that were used to train the generator. Practically, the FID score is calculated by the L2 distance of the mean and variance of the image feature map distributions extracted by pretrained Inception V3 [44]. A smaller FID score indicates better performance.

Baselines: We compare our method with two representative image-to-image translation models, pix2pixHD and SPADE. pix2pixHD is a pix2pix-based conditional image synthesis framework for high-resolution. SPADE is another conditional image synthesis framework with normalization layer that encode spatial semantic information of condition image. We also construct two ablation models of our network, w/o IAN and w/o Shape Loss. In the former model, IAN is replaced by SPADE normalization layer. In the latter model, we remove the shape enhancing loss from model training.

4.2 Comparison results

Qualitative Results: Figure 5 compares the images generated by the proposed model, the existing models, and the ablation models. In the figure, we can observe that the images by the

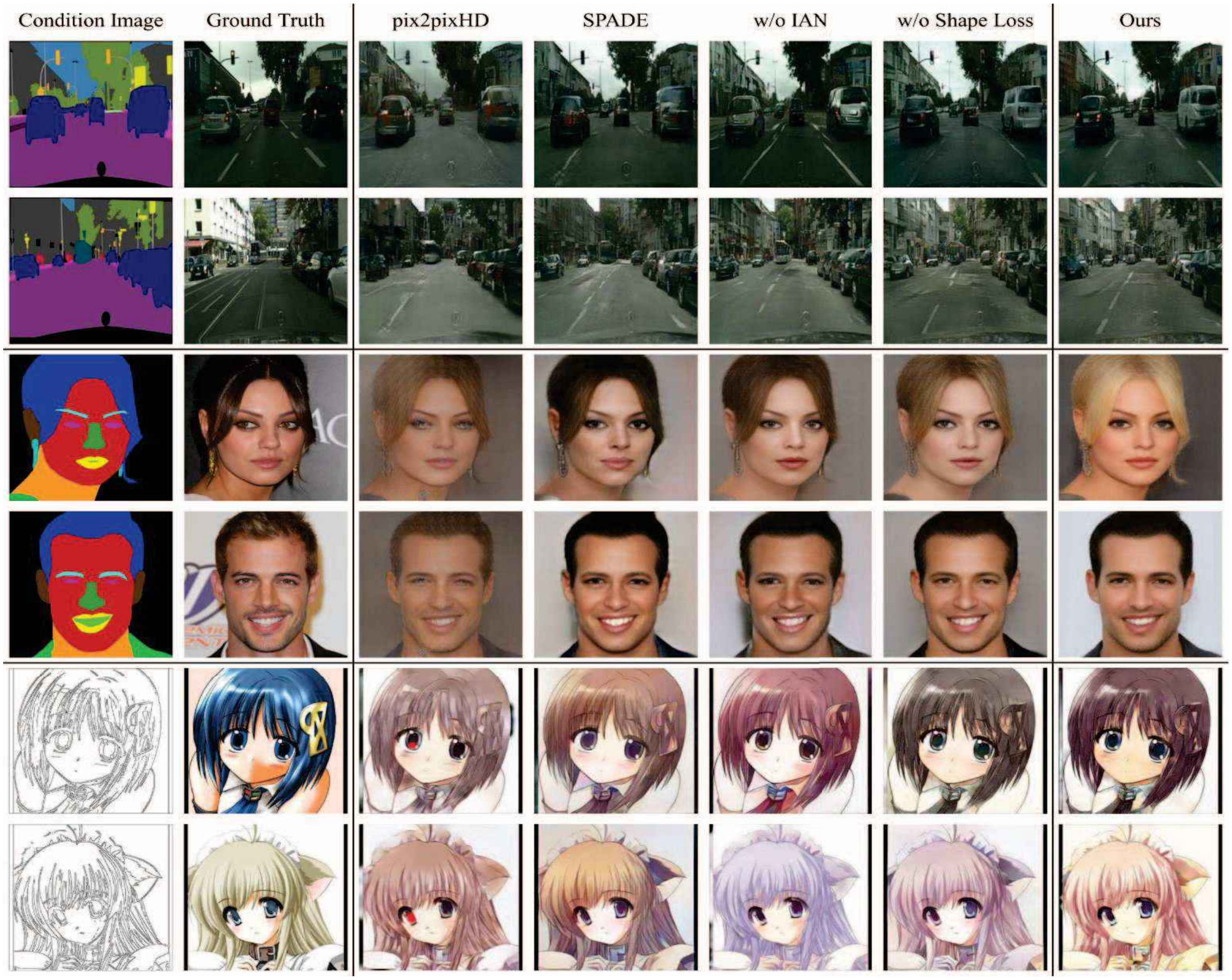

Fig. 5. Generated image comparison with existing methods and ablation models. Our methods generate more realistic images with diverse color expression and vivid shapes. Rows 1~2, 3~4, 5~6 are the results of Cityscape, CelebAMask-HQ, High-Resolution Anime Face dataset, respectively. (Please zoom for detail look)

proposed model more effectively describe the style quality and shape according to the domain. That is, the proposed model reproduces the local style of a specific segmentation label while maintaining the global style according to the domain. Especially in the anime dataset, it can be difficult to express an appropriate style because the input image has only line information excluding spatial semantic labels. However, IAN effectively generated color expression of appropriate animation styles according to regions such as eyes, hair, and skin. In contrast, existing models and w/o IAN produced faded color. Fig. 6 shows the binary edge maps of the generated images to evaluate shape representation. Our model produced sharper and more realistic edge maps compared to the existing models. In addition, the results of the ablation model w/o Shape Loss indicate that the shape enhancing loss contributes to shape representation. Overall, the proposed IAN and shape enhancement loss effectively perform style representation and shape representation. Fig. 7 provides more examples.

TABLE I. FID scores of our methods and comparative model. FID scores were measured using validation sets and same number of generated images. ‘CelebA’ and ‘Anime' denote CelebAMask-HQ and High-Resolution Anime Face dataset.

Dataset	pix2pixHD	SPADE	w/o IAN	w/o Shape Loss	Ours
Cityscape	74.57	71.04	66.21	60.46	**59.49**
CelebA	107.85	104.77	101.29	97.51	**95.55**
Anime	113.29	105.92	102.06	96.00	**93.41**

Quantitative Results: Now, we evaluate the quality of the generated image using the FID score. As shown in Table. 1, our model outperforms the existing and ablation models for all datasets. From the ablation models, we can see that IAN and shape enhancing loss are combined together to achieve the best image quality without other adverse effects. This is probably because the two methods are designed to optimize different objects (style and shape representation), with minimal influence on each other.

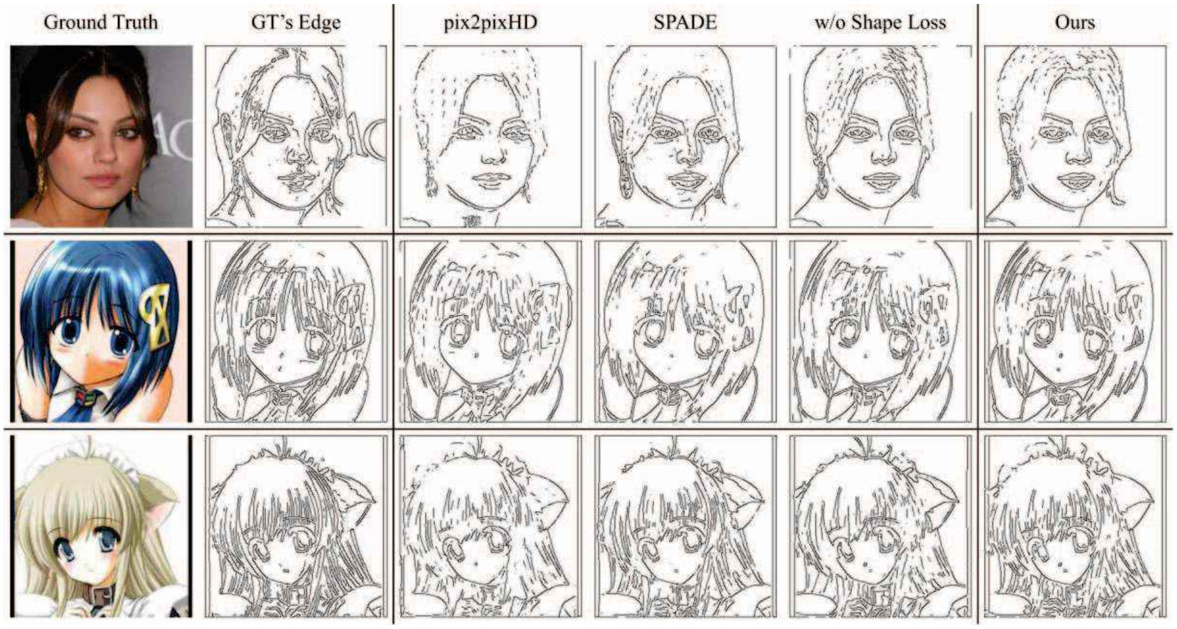

Fig. 6. Binary edge maps of generated images with existing methods and ablation models. Our method results realistic and distinct edge maps compared to other methods. First, second and third rows are the results of Cityscape, CelebAMask-HQ, High-Resolution Anime Face dataset. (Please zoom for detail look)

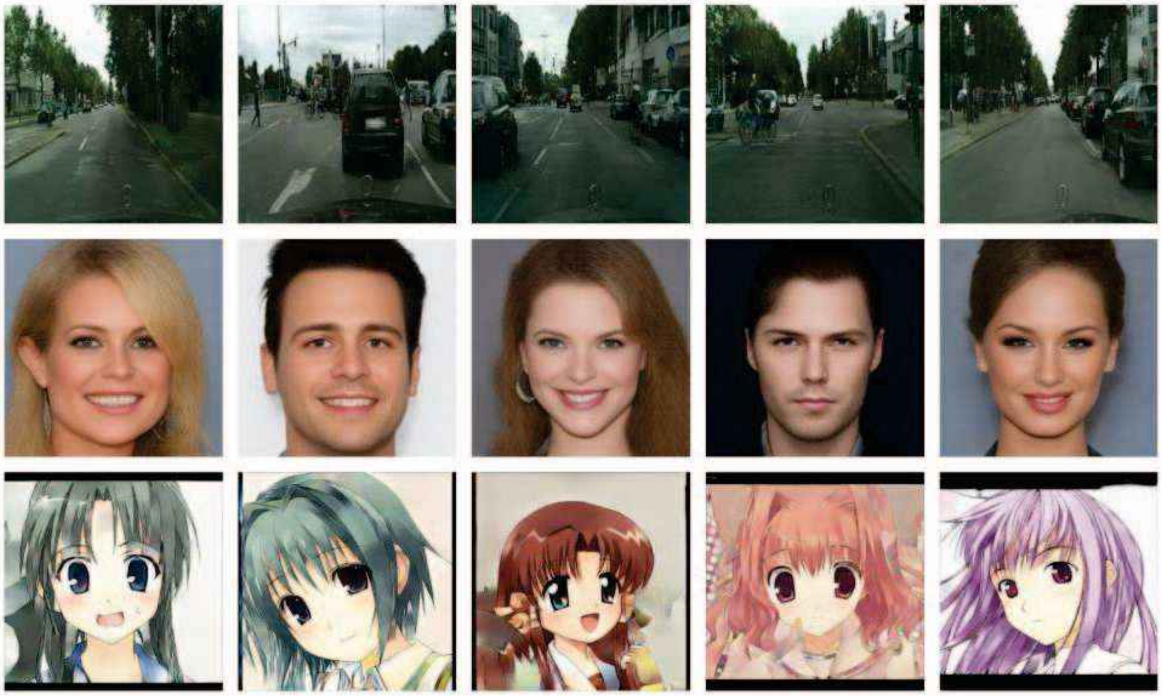

Fig. 7. Additional results of our models. First, second and third rows are the results of Cityscape, CelebAMask-HQ, High-Resolution Anime Face dataset.

V. Conclusion

In this paper, we proposed a novel GAN-based conditional image synthesis model that improved style representation and shape representations. In particular, we enhanced global and local style representations using a conditional normalization layer called IAN, and improved the shape of generated images using a shape enhancing loss. In comparative experiments with two popular image-to-image transformation models on three popular datasets, the proposed model performed better in terms of style and shape representation. We also showed in the ablation study that IAN and the shape enhancing loss are combined together to achieve the best image quality without interfering with each other.

References

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," ACM SIGGRAPH, 2009

[2] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” In International Conference on Learning Representations, 2019

[3] Q. Chen and V. Koltun, “Photographic image synthesis with cascaded refinement networks,” In IEEE International Conference on Computer Vision, 2017

[4] K. Nazeri, E. Ng, and M. Ebrahimi, "Image colorization using generative adversarial networks," International conference on articulated motion and deformable objects , Springer, 2018.

[5] Y. Jing, et al., "Neural style transfer: A review." IEEE transactions on visualization and computer graphics vol. 26, no. 11, pp.3365-3385, 2019.

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

[7] T. Park, M. Liu, T. Wang and J. Zhu, "Semantic image synthesis with spatially-adaptive normalization," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019.

[8] H. Vries, et al., "Modulating early visual processing by language," Advances in Neural Information Processing Systems, 2017.

[9] G. Seif and D, Androutsos, "Edge-based loss function for single image super-resolution," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, 2018.

[10] K. Jiang, et al., "Edge-enhanced GAN for remote sensing image superresolution," IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 8, pp. 5799-5812, 2019.

[11] J. Johnson, A, Alahi and L Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," European conference on computer vision, 2016.

[12] I. Goodfellow, et al., "Generative adversarial networks," Communications of the ACM, vol. 63, no.11, pp. 139-144, 2020.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint, 2013.

[14] M. Mirza, and S. Osindero, "Conditional generative adversarial nets," arXiv preprint, 2014.

[15] R. Gal, et al., "Stylegan-nada: Clip-guided domain adaptation of image generators," arXiv preprint, 2021.

[16] S. Reed, et al., "Generative adversarial text to image synthesis," International conference on machine learning, 2016.

[17] Y. Choi, et al., "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

[18] X. Huang, M. Liu, S. Belongie and J. Kauts, "Multimodal unsupervised image-to-image translation," Proceedings of the European conference on computer vision, 2018.

[19] S. Loffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," International conference on machine learning, 2015.

[20] D. Ulyanov, A. Vedaldi and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," arXiv preprint, 2016.

[21] J. L. Ba, J. R. Kiros and G. E. Hinton, "Layer normalization," arXiv preprint, 2016.

[22] Y. Wu, K. He, "Group normalization," Proceedings of the European conference on computer vision, 2018.

[23] X. Wang, K. Yu, C. Dong and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint, 2014.

[25] A. Lucas, S. Lopez-Tapia, R. Molina and A. K. Katsagelos, "Generative adversarial networks and perceptual losses for video super-resolution," IEEE Transactions on Image Processing, vol. 28, no. 7, pp. 3312-3327, 2019.

[26] D. Wolr, J. Prankl and M. Vincze, "Fast semantic segmentation of 3D point clouds using a dense CRF with learned parameters," 2015 IEEE International conference on robotics and automation, 2015.

[27] G. Yang, Q. Zhang and G. Zhang, "EANet: Edge-aware network for the extraction of buildings from aerial images," Remote Sensing, vol. 12, no. 13, pp. 2161, 2020.

[28] S. Ko and B. Dai, "Multi-laplacian GAN with edge enhancement for face super resolution," 2020 25th International Conference on Pattern Recognition, 2021.

[29] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. C. Courville, "Improved training of wasserstein gans," Advances in neural information processing systems 30, 2017.

[30] A. Radford, L. Metz and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint, 2015.

[31] L. Ding and A. Goshtasby, "On the Canny edge detector," Pattern recognition vol. 34, no. 3, pp. 721-725, 2001.

[32] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," International Conference on Medical image computing and computer-assisted intervention, 2015.

[33] K. He, X..Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[34] T. Wang, et al., "High-resolution image synthesis and semantic manipulation with conditional gans," Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

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

[36] X. Mao, et al., "Least squares generative adversarial networks," Proceedings of the IEEE international conference on computer vision, 2017.

[37] 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 30, 2017.

[38] D. P. Kingma and J.Ba, "Adam: A method for stochastic optimization," arXiv preprint, 2014.

[39] M. Cordts, et al., "The cityscapes dataset for semantic urban scene understanding," Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[40] C. Lee, Z. Liu, L. Wu and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[41] T, Karras, et al., "Progressive growing of gans for improved quality, stability, and variation" arXiv preprint, 2017.

[42] B. Gwern, Anonymous and The Danbooru Community, “Danbooru2019 Portraits: A Large-Scale Anime Head Illustration Dataset,” Web, 2019

[43] T. Salimans, et al., "Improved techniques for training gans," Advances in neural information processing systems 29, 2016.

[44] C. Szegedy, et al., "Rethinking the inception architecture for computer vision," Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.