Bulletin of Networking, Computing, Systems, and Software – www.bncss.org, ISSN 2186-5140
Volume 14, Number 1, pages 81–84, January 2025

Facial Sketch Synthesis using Attention guided CycleGAN with Surface Normal Loss

Kexuan Yang, Victor Parque, Koji Nakano, and Yasuaki Ito
Graduate School of Advanced Science and Engineering, Hiroshima University
Kagamiyama 1-4-1, Higashi-hiroshima, 739-8527, JAPAN

***Abstract*—Sketches play an important role in encoding representative facial features through simple strokes. The state-of-the-art approaches for sketch synthesis using generative and unsupervised learning methods have been able to generate perceptual facial sketches. However, such methods are prone to render low-quality facial sketches due to the effect of the geometry of faces and the portrait background. In this paper, to tackle this problem, we formulate and incorporate new loss functions based on facial surface normals and edge maps into a generative adversarial training framework based on CycleGAN. Facial surface normals are estimated by a convolutional neural network based on Dense Inductive Biases for Surface Normal Estimation (DSINE) and encoded into colored frames indicating normal directions of pictures, better capturing the geometric information of the face. Edge maps are estimated by Holistically-Nested Edge Detection (HED), better capturing the lines and contours of face pictures. Furthermore, we extend the attention-guided generator that separates foreground and background during training, thereby reducing the impact of background elements on the face sketch. As such, our approach aims to learn generator architectures to translate pictures to sketches and vice versa with utmost consistency and geometric accuracy. Our computational experiments using the FS2K (containing annotated facial 2,104 images) on an Nvidia A6000 show the improved performance in terms of Learned Perceptual Image Patch Similarity (LPIPS, 0.412), Structural Similarity (SSIM, 0.456), Multiscale Structural Similarity (MS-SSIM, 0.515), Feature Similarity (FSIM, 0.609), Structure Co-Occurrence Texture (SCOOT, 0.404). Our results have the potential to study improved cycle-consistent architectures to generate face sketches with high-quality and rich details.**

***Index Terms*—Generative Adversarial Networks, Facial Sketch Synthesis (FSS), Deep Learning, Surface Normal**

I. Introduction

Sketching is a fundamental technique in many forms of art and design, ans is often viewed as the first step in the creative process. By using simple lines, an artist can outline the overall structure of the subject, which can be the foundation for further refinement and detailing. In the area of computer vision, sketching can be viewed as an image-to-image translation task, which aims to learn a mapping from one image domain to another one. For example, as shown in Figure 1, the model translates an image (e.g., a photo) into another one (e.g., a sketch). In recent years, the advancement of deep learning models, particularly of Generative Adversarial Networks (GANs) [1], have triggered new architectures for image-to-image translation 1 tasks due to the capacity to learn complex data distributions and the feasibility to generate high-quality images. In this paper, we utilize deep learning methods for face sketch synthesis and aim to improve the quality of generated face sketches by enhancing line accuracy, geometric structure, and reducing background noise.

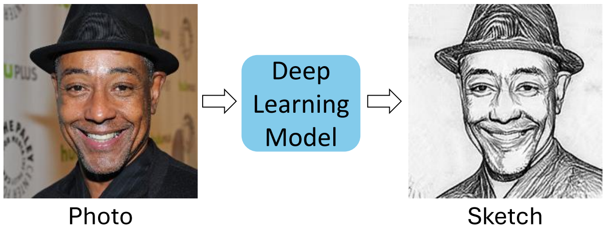

Fig. 1. Sketching as Image-to-Image translation

For sketches, lines and geometric structures are of utmost importance. A well-drawn sketch can capture the essence of an object or scene using just a few lines, making it crucial to preserve the integrity of the underlying structure during the translation process. Surface normals encode geometric information into RGB colors [2], allowing us to leverage this representation to capture the geometric details of objects. Here we propose a method for face sketch synthesis that uses both edge maps and surface normal maps to ensure the accurate representation of facial lines and geometric structures. Our model takes facial photos as input to generate face sketches, which are then reconstructed back into face photos. Edge maps and surface normal maps are computed for both the input face photos and the reconstructed face photos, and during training, the model is optimized to minimize the differences between these pairs. By aligning both edge and surface normal maps, we aim to improve the geometric fidelity of the generated sketches, resulting in more accurate representations.

One of the key challenges in face sketch synthesis is minimizing the influence of background elements. Many face-related datasets contain images with various background components, which can introduce noise and affect the accuracy of the model's output. We draw inspiration from Attention-GAN [3] to enhance the model's focus on the facial region during the sketch generation process. In our model, we utilize the generator from AttentionGAN to separate foreground and background, enabling each to be processed independently during training. This separation allows us to reduce the influence of background elements on the face sketch, resulting in cleaner and more accurate sketches that focus primarily on the facial features.

In this paper, our contributions are as follows:

- We propose an approach that uses edge maps and surface normal maps to enhance the geometric accuracy of face sketches. This combination allows for both the preservation of prominent facial outlines and the accurate details of facial geometry.
- Second, we use an attention mechanism based on AttentionGAN to focus on the facial region. This reduces the impact of background elements and ensures that the generated sketches emphasize the essential features of the face.
- We trained our model on the FS2K dataset and evaluated it alongside existing models using LPIPS, SSIM, MS-SSIM, FSIM, and SCOOT metrics. Our model achieved competitive results in LPIPS, SSIM, and MS-SSIM, and outperformed others in FSIM and SCOOT.

In the rest of the paper, we will explain the specific implementation details and present the results of the related experiments.

II. Related Works

A. Image-to-Image Translation

In the field of image-to-image translation, one of the most well-known frameworks is Pix2Pix [4], which uses a conditional GAN to learn a mapping from paired images in different domains. The supervised learning approach in Pix2Pix has been widely applied in tasks such as sketch-to-photo translation, style transfer, and semantic segmentation. Pix2Pix demonstrated that GANs could successfully generate high-quality image transformations when paired datasets are available, yet the reliance on paired training data limits its applicability in certain cases.

To address the limitations of paired datasets, researchers have proposed alternative methods. For instance, CycleGAN [5] introduced an unsupervised approach to learn the mapping between two domains without the need for paired samples by using cycle consistency loss enabling the translation between domains where obtaining paired datasets is difficult or impractical, such as photo-to-sketch or artistic style transfer. And our model also based on CycleGAN framework. However, despite the above-mentioned merits, CycleGAN's unsupervised nature can sometimes result in lower geometric fidelity, which is critical in tasks like face sketch synthesis.

B. Face Sketch Synthesis

For face sketch synthesis task, preserving the details of facial features while maintaining overall geometric accuracy is challenging. There are currently many deep learning models for face sketch synthesis, such as APDrawingGAN [6], which combines global and local networks to capture both the overall structure of the face and specific facial features. UPDG [7] addresses face sketch synthesis by introducing an asymmetric cycle mapping, which ensures visible reconstruction information is embedded selectively in facial regions. With localized discriminators for facial features and a style classifier to manage multiple drawing styles, the model preserves key facial details and generates face sketches in a variety of styles. Learning-to-draw [8] proposed a GAN-based framework with additional geometric constraints, incorporating depth maps to ensure the structure is accurately captured in the generated sketches. Additionally, the CLIP model is also used to ensure semantic consistency in the generated images.

In our work, we extended the CycleGAN-based facial sketch synthesis by using edge maps and surface normal maps to better capture the contours and shapes in the generated facial sketches. Unlike existing approaches, our model utilizes surface normal maps to encode relevant geometric information, further improving the realism of the generated sketches.

III. Proposed Method

In this section, we will introduce the detailed implementation of the model.

A. Model

The goal of our research is to train a model that can convert face photographs into corresponding sketches. To achieve this goal, we use a dataset containing face photographs and their associated sketches. We treat this problem as an unpaired image-to-image translation task between two domains: domain A, which consists of photographs, and domain B, which consists of sketches. Since the image pairs are unaligned, preserving the overall structural integrity between the two domains is critical. Therefore, we adopt a CycleGAN-based architecture [5], leveraging cycle consistency to ensure that the translations between the two domains remain coherent and faithful to the original inputs.

The structure of the model is shown in Figure 2. Our model employs two generator networks, G_A and G_B, responsible for translating images from domain A to domain B and vice versa. Additionally, two discriminator networks, D_A and D_B, are used to differentiate between real images in each domain and the generated images from the opposing domain. This adversarial training setup ensures that the generated sketches not only resemble real sketches stylistically but also preserve the structure of the original photographs.

To further ensure the accuracy of both the lines and geometric structure of the generated face sketches, our model incorporates edge and geometric information to guide the translation process. Specifically, we utilize pre-trained HED (Holistically-Nested Edge Detection) [9] and DSINE [2] models, which have demonstrated strong performance in edge detection and surface normal estimation, respectively. In order to guide the model training using the aforementioned models, we introduce the edge loss and surface normal loss, which will be specifically discussed in Section 3.2.

Since the images in the dataset contain backgrounds, we improved the generator structure to reduce the impact of the image background on training, drawing inspiration from the architecture of AttentionGAN [3]. The structure of the generator is shown in Figure 3. The most important components of the generator are the Attention Mask Generator and the Content Mask Generator. The Attention Mask Generator aims to produce both foreground and background attention

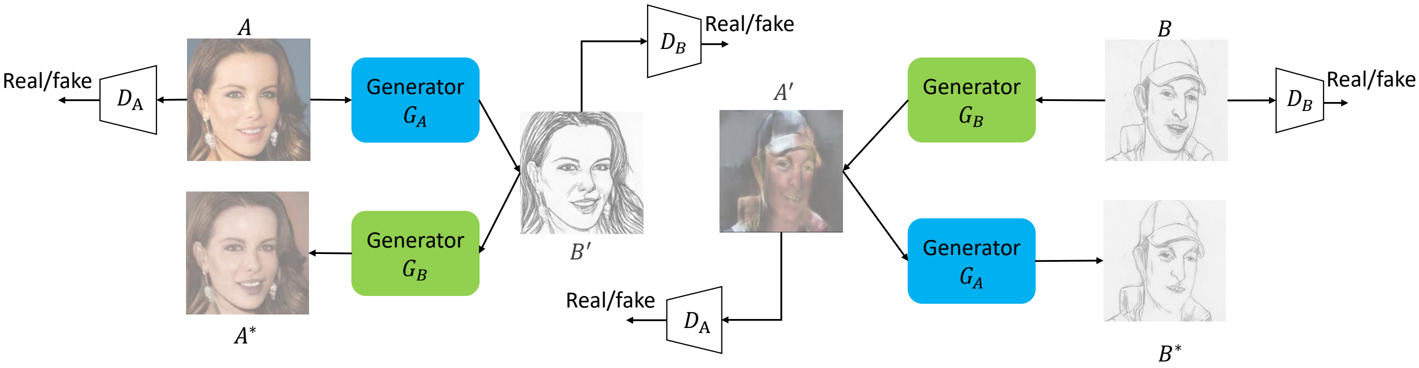

Fig. 2. CycleGAN-based architecture

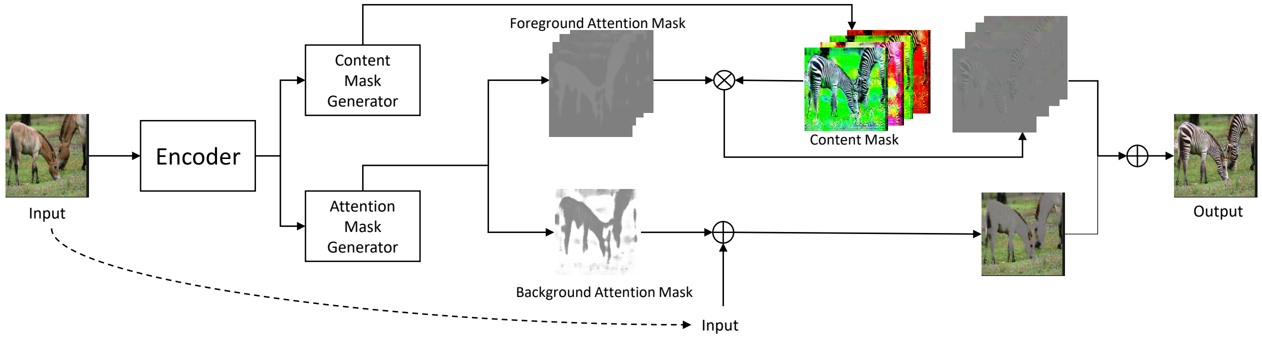

Fig. 3. Generator structure

masks, which are used to selectively focus on the relevant content from the corresponding content masks generated by the Content Mask Generator and the input image.

Using the Attention Mask Generator and the Content Mask Generator illustrated in Figure 3, we generate foreground attention masks, a background attention mask and content masks. These attention masks are applied to selectively focus on different regions of the input image, allowing us to process the foreground (e.g., the face) and the background separately. The Content Mask Generator further produces content masks that represent the useful information in these regions. By combining the attention masks with the content masks, we can effectively extract relevant features from both the face and the background while minimizing the influence of background elements on the face during training. This separation ensures that the model learns to prioritize facial features without being distracted by irrelevant background details.

By using a CycleGAN-based framework for unpaired image translation with additional supervision from pre-trained models in edge detection and surface normal estimation, our approach ensures that the generated face sketches are both stylistically accurate and structurally faithful to the original face photographs. Additionally, we improved the generator by leveraging the architecture of AttentionGAN, minimizing the influence of the background on the face during training as much as possible.

B. Loss Functions

The adversarial loss

In the face sketch synthesis task, the adversarial loss plays a critical role in guiding the generator to produce sketches. By employing a generator-discriminator setup, the generator attempts to create sketches from input images, while the discriminator learns to distinguish between real sketches and generated ones. The adversarial loss encourages the generator to improve the quality of the sketches by minimizing the discriminator's ability to distinguish between the two.

The adversarial loss for the generator G and the discriminator D can be defined as:

$$\begin{aligned} L_{\text{GAN}} = \mathbb{E}_{a\sim A}\left[D_A(a)^2\right] + \mathbb{E}_{b\sim B}\left[(1 - D_A(G_B(b)))^2\right] \\ + \mathbb{E}_{b\sim B}\left[D_B(b)^2\right] + \mathbb{E}_{a\sim A}\left[(1 - D_B(G_A(a)))^2\right] \end{aligned} \tag{1}$$

where a represents the real photographs from domain A, and b represents the real sketches from domain B; $G_A(a)$ represents the generated sketches from photographs, and $G_B(b)$ represents the generated photographs from sketches; D_A and D_B are the outputs of the discriminators for real and generated images in their respective domains. The first two terms ensure that the generated sketches resemble real sketches, and the last two terms ensure that the generated photographs resemble real photographs.

The cycle consistency

In face sketch synthesis tasks, cycle consistency loss ensures that the mapping between the photograph domain and the sketch domain is consistent. It enforces that if a photograph is translated into a sketch and then back into a photograph, the result should closely match the original photograph. This is important for maintaining the structural integrity of facial features during translation. The cycle consistency loss is defined as follows:

$$L_{\text{cycle}} = \|G_B(G_A(a)) - a\| + \|G_A(G_B(b)) - b\| \tag{2}$$

where G_A and G_B are the same as those introduced in the adversarial loss section; the first term $\|G_B(G_A(a)) - a\|$ ensures that converting a photograph to a sketch and then back to a photograph results in the original photograph; and the second term $\|G_A(G_B(b)) - b\|$ ensures the same for sketches, maintaining the cycle consistency in both directions.

The surface normal loss

The task of surface normal estimation involves predicting the orientation of surfaces in an image, which is represented through a normal map. A surface normal map encodes the 3D orientation of surfaces by representing the surface normals as RGB values, where the red, green, and blue channels correspond to the X, Y, and Z components of the normal vector, respectively. This encoding is crucial for preserving the geometric structure of facial features in face sketch synthesis.

In our model, we incorporate the state-of-the-art DSINE model for surface normal estimation to guide the generation process. The role of the DSINE model in our setup is similar to that of cycle consistency. It ensures that the geometric structure of the generated sketch closely aligns with the original input image. The surface normal loss is calculated in the model as shown in Figure 4. The surface normal loss is formulated as follows:

$$L_{\text{normal}} = \frac{1}{N}\sum_{i=1}^{N}(\text{DSINE}(a)_i - \text{DSINE}(G_B(G_A(a)))_i)^2 \tag{3}$$

where a represents the input photograph, and $G_B(G_A(a))$ represents the photograph that is generated by first translating the input into a sketch using G_A, and then translating the sketch back into a photograph using G_B, similar to the cycle consistency approach. The surface normal loss L_{normal} is computed using the Mean Squared Error (MSE) loss between the surface normals estimated by the DSINE model for both the input photograph a and the reconstructed photograph $G_B(G_A(a))$. Specifically, N denotes the total number of pixels, and i represents the index of each pixel. The goal is to minimize the difference between the surface normals of the original and reconstructed images, ensuring consistency in geometry through the transformations.

The edge loss

Edge maps can highlight the important structural lines of an image. In our model, inspired by [7], we use the HED (Holistically-Nested Edge Detection) [9] model to detect edges in both the input and the reconstructed input. The edge maps are then used to guide the training process, ensuring that the generated sketch preserves the key structural lines of the input image.

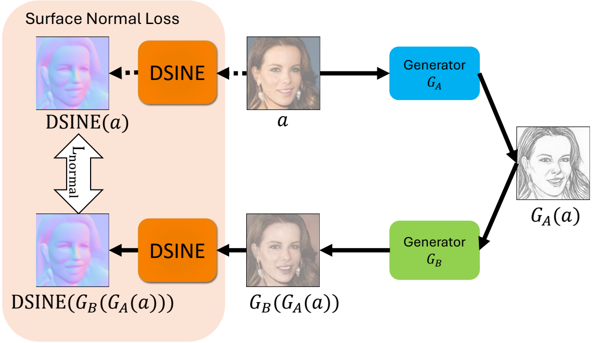

Fig. 4. Surface Normal Loss

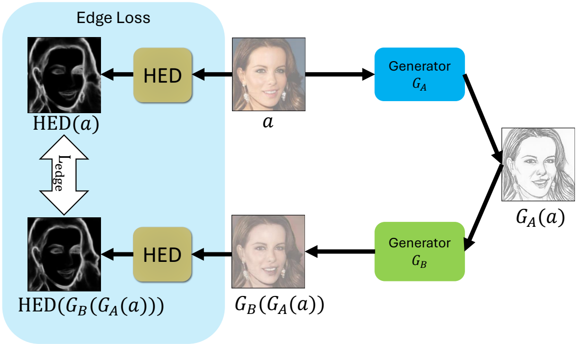

Fig. 5. Edge Loss

The edge loss also functions in conjunction with the cycle consistency to maintain the fidelity of the edges throughout the translation process. The loss is defined using the LPIPS (Learned Perceptual Image Patch Similarity) [10] metric, which measures the perceptual similarity between two images. The edge loss is calculated in the model as shown in Figure 5. The edge loss is given by:

$$L_{\text{edge}} = \text{LPIPS}(\text{HED}(a), \text{HED}(G_B(G_A(a)))) \tag{4}$$

In this equation, L_{edge} represents the edge loss, calculated using the LPIPS function. The HED model estimates the edges for both the input photograph a and the reconstructed photograph $G_B(G_A(a))$, and the LPIPS function measures the perceptual similarity between these two edge maps.

Full losses The overall loss function is as follows:

$$\begin{aligned} L = {} & \lambda_{\text{edge}} L_{\text{edge}} + \lambda_{\text{normal}} L_{\text{normal}} \\ & + \lambda_{\text{GAN}} L_{\text{GAN}} + \lambda_{\text{cycle}} L_{\text{cycle}} \end{aligned} \tag{5}$$

The parameters were configured as $\lambda_{\text{edge}} = 1$, $\lambda_{\text{normal}} = 10$, $\lambda_{\text{GAN}} = 1$, $\lambda_{\text{cycle}} = 0.1$.

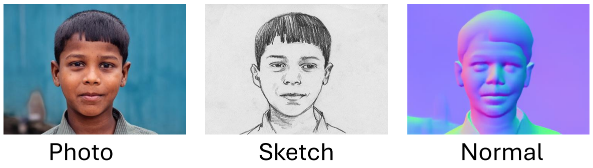

Fig. 6. Samples from dataset

IV. Experiments

A. Dataset

For our experiments, we utilized the FS2K dataset, a high-quality face sketch synthesis dataset introduced in the FSGAN [11] paper. The FS2K dataset consists of 2,104 image-sketch pairs, covering a broad spectrum of variations, including diverse image backgrounds, skin tones, sketching styles, and lighting conditions. The variety ensures that the dataset provides a comprehensive and robust foundation for training deep learning models in face sketch synthesis tasks.

Since our model requires the use of surface normal maps to guide the training process, we added pseudo-ground truth surface normal maps to the existing FS2K dataset to ensure training efficiency and stability. Specifically, the facial photographs used for training are processed using the DSINE model for surface normal estimation, and the generated normal maps are saved in the dataset. The file names of the generated normal maps must match the original facial photograph file names to ensure paired usage during training. An example of the processed dataset is shown in Figure 6.

B. Setup and Results

The model was trained using PyTorch [12] under the following experimental settings: 150 epochs, a batch size of 4, and a learning rate of 2.0×10^{-4}. The Adam [13] optimizer was used to optimize the network. All experiments were conducted on an Nvidia A6000 GPU. The results obtained from this training configuration are shown in Figure 8. It is evident that our results not only maintain a close resemblance to the original portraits but also exhibit the characteristics of hand-drawn sketches.

To compare the performance of different models on the FS2K dataset's test set, we conducted a quantitative comparison, comparing our model with CycleGAN [5], Learn2Draw [8] and FSGAN [11], using LPIPS [10], SSIM [14], MS-SSIM [15], FSIM [16] and SCOOT [17] metrics. In Table I, the values highlighted in red indicate the best performance, while underlined values represent the second-best results. Our proposed model achieves the second-best performance in LPIPS, SSIM, and MS-SSIM, demonstrating competitive results in terms of perceptual similarity and structural preservation. Furthermore, our model achieves the best scores in both FSIM and SCOOT. Notably, the SCOOT metric is specifically designed for evaluating face sketches, indicating that (1) our method produces results that are closer to hand-drawn face sketche, and (2) our approach is highly effective in capturing the fine details and stylistic nuances characteristic of human-drawn facial sketches, making it well-suited for the face sketch synthesis task.

TABLE I
Quantitative comparison between models using LPIPS, SSIM, MS-SSIM, FSIM and SCOOT metrics.

	LPIPS↓	SSIM↑	MS-SSIM↑	FSIM↑	SCOOT↑
CycleGAN	0.462	0.364	0.419	0.563	0.331
Learn2Draw	0.429	0.408	0.473	0.583	0.392
FSGAN	0.369	0.508	0.583	0.490	0.275
Ours	0.412	0.456	0.515	0.609	0.404

However, the aforementioned metrics do not always fully reflect the performance of models in the facial sketch synthesis task, as deformations are inevitable during sketch drawing. This leads to discrepancies between the hand-drawn sketches in the dataset and the real photos (such as differences in the direction of the eyes or the presence of glasses). Therefore, we also need to conduct subjective evaluations. As shown in Figure 9, we compared the three models, and it can be observed that the face sketches generated by our model are closer to the hand-drawn sketches in the ground truth. In contrast, the results from Learn2Draw and CycleGAN appear somewhat messy and contain noise. Notably, although FSGAN performs well on the metrics in Table I, the generated face sketches exhibit distortions and deformations, failing to accurately reflect the original facial structure.

We also conducted an ablation study on the edge and surface normal modules. The corresponding results are shown in Table II and Figure 10. In Table II, By comparing with models where certain modules were removed, our full model achieved second place in the LPIPS, SSIM, and FSIM metrics. In the face-sketch-specific SCOOT metric, the full model still attained first place. In FIgure 10, it can be seen that after removing these modules, the generated face sketches became blurry, noisy, or exhibited disorganized lines.

TABLE II
Results of ablation studies.

	LPIPS↓	SSIM↑	MS-SSIM↑	FSIM↑	SCOOT↑
w/o Edge & Normal	0.478	0.400	0.520	0.609	0.392
w/o Normal	0.401	0.465	0.519	0.608	0.380
w/o Edge	0.454	0.421	0.526	0.614	0.379
Full	0.412	0.456	0.515	0.609	0.404

V. Limition and Future Work

We further evaluated the model on non-facial images. While our model is capable of generating corresponding sketches, there is a noticeable loss of detail. As illustrated in Figure 7, elements such as animal fur, brick textures, tree branches, and architectural details are partially lost in the generated sketches.

During the training process, we utilized pseudo-ground truth and a pre-trained model to generate surface normals. The accuracy of the generated surface normals may also have an impact on the final results.

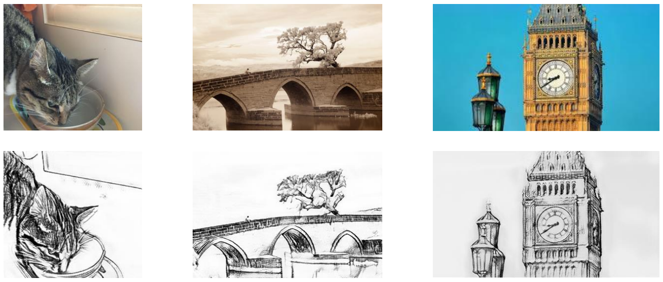

Fig. 7. Test on non-facial images

In future work, we will explore the performance of our model on additional datasets and continue to focus on improving the quality of the generated sketches.

VI. Conclusion

In this paper, we presented an approach to face sketch synthesis, dealing with problems related to facial geometry and background noise by using edge maps and surface normal maps into a CycleGAN-based architecture. This method improves geometric fidelity and structural accuracy by guiding the model to focus on key facial features while minimizing irrelevant background influence. In future work, we will focus on providing users with more control over the generated face sketches and further improving the quality of the generated images.

References

[1] 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.

[2] G. Bae and A. J. Davison, "Rethinking inductive biases for surface normal estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9535–9545.

[3] H. Tang, H. Liu, D. Xu, P. H. Torr, and N. Sebe, "AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks," *IEEE transactions on neural networks and learning systems*, vol. 34, no. 4, pp. 1972–1987, 2021.

[4] 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.

[5] 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.

[6] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical gans," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 743–10 752.

[7] ——, "Unpaired portrait drawing generation via asymmetric cycle mapping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8217–8225.

[8] C. Chan, F. Durand, and P. Isola, "Learning to generate line drawings that convey geometry and semantics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7915–7925.

[9] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[10] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[11] D.-P. Fan, Z. Huang, P. Zheng, H. Liu, X. Qin, and L. Van Gool, "Facial-sketch synthesis: A new challenge," *Machine Intelligence Research*, vol. 19, no. 4, pp. 257–287, 2022.

[12] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[14] 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.

[15] 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.

[16] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[17] D.-P. Fan, S. Zhang, Y.-H. Wu, Y. Liu, M.-M. Cheng, B. Ren, P. L. Rosin, and R. Ji, "Scoot: A perceptual metric for facial sketches," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5612–5622.

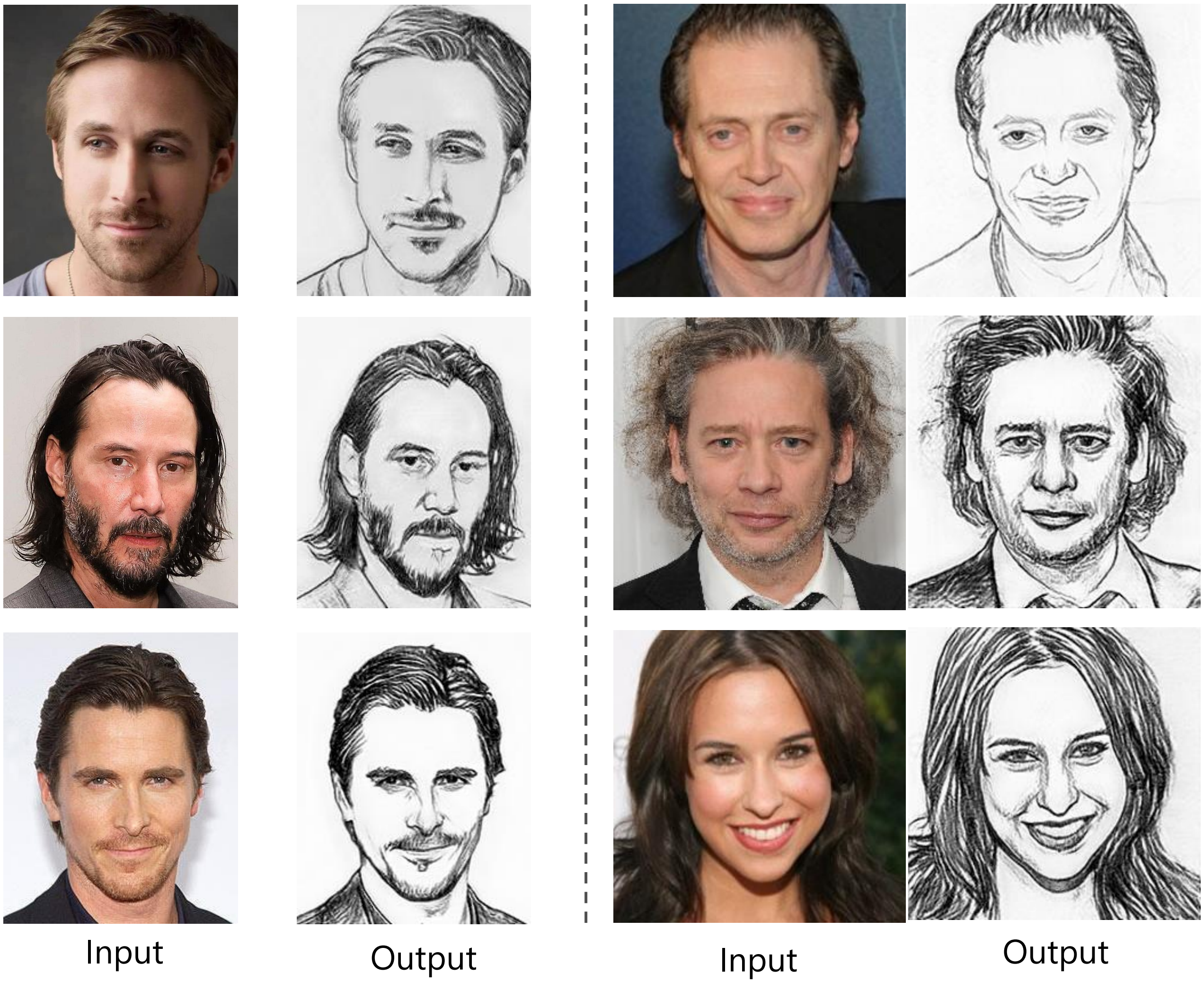

Fig. 8. Results generated by our model

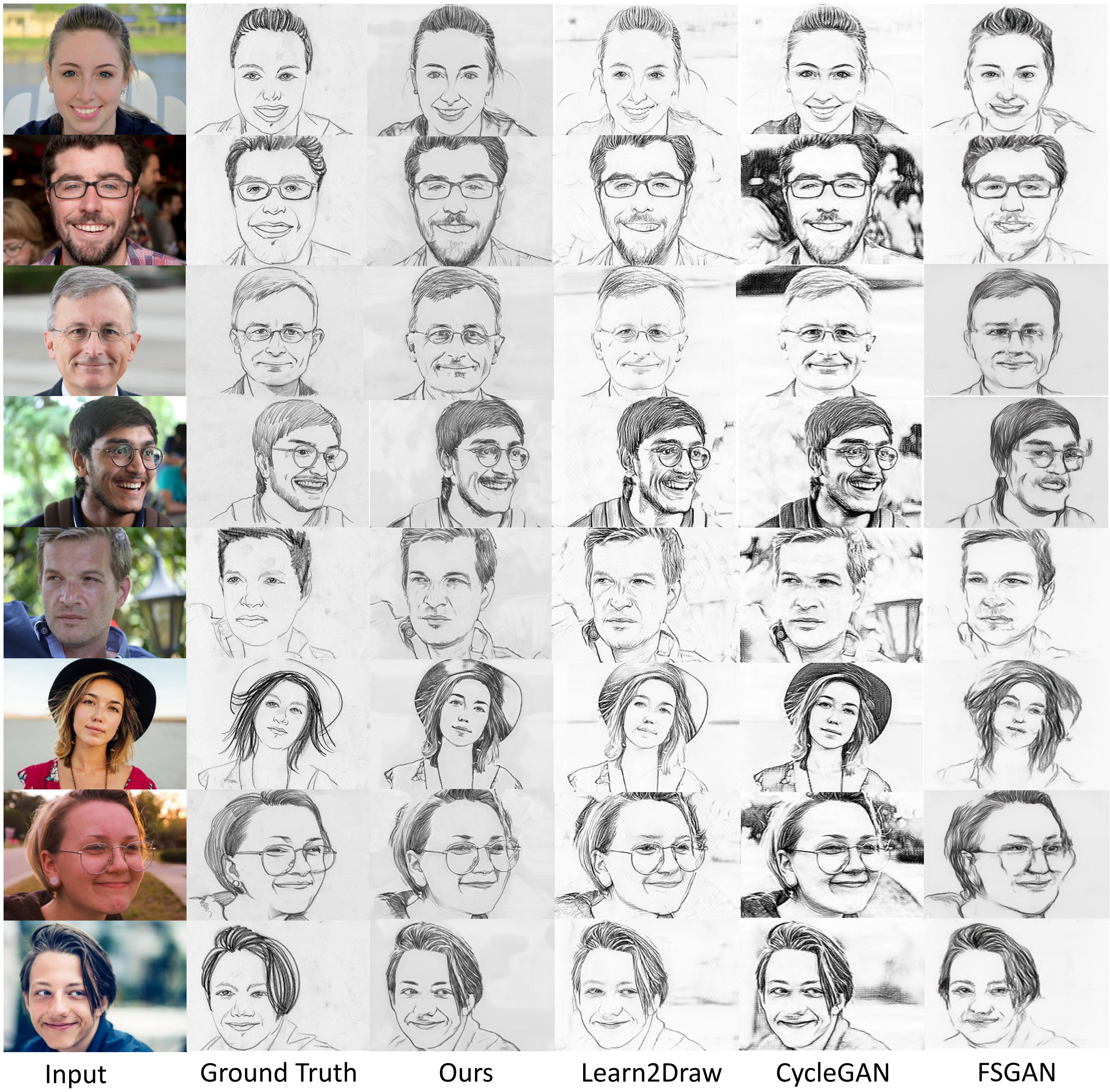

Fig. 9. Compare with CycleGAN, Learn2Draw and FSGAN

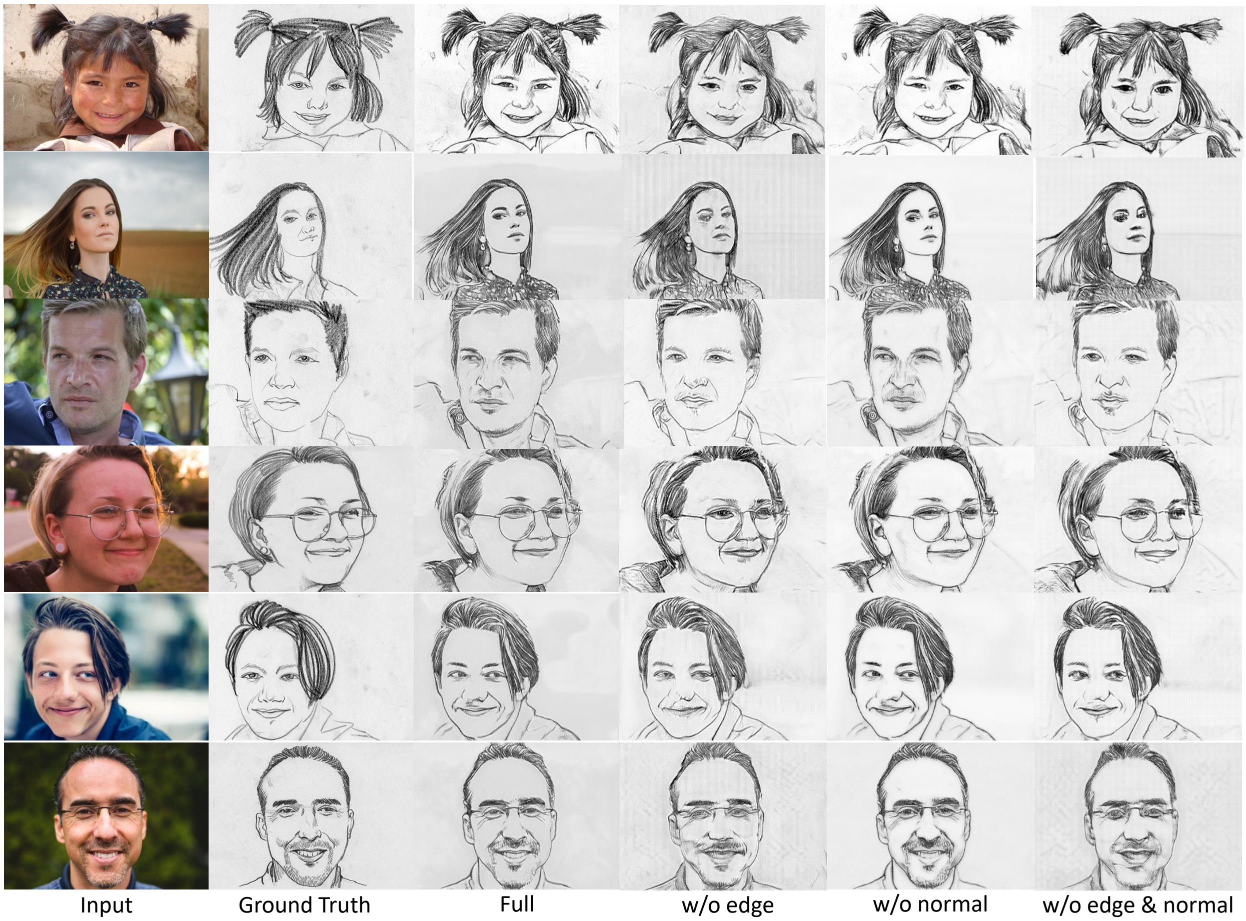

Fig. 10. Ablation study