CONDITIONAL DIFFUSION INPAINTING FOR SKETCH-TO-FACE SYNTHESIS

Sanhita Pathak*[1], Vinay Kaushik[2], and Brejesh Lall[1]

[1]Indian Institute of Technology Delhi
[2]Indian Institute Of Information Technology Sonepat
Sanhita.Pathak@dbst.iitd.ac.in, vkaushik@iiitsonepat.ac.in, brejesh@ee.iitd.ac.in

Abstract

Generating realistic faces from monochromatic sketches is challenging due to missing details like expressions and skin tones. GANs struggle with stability and structure, while diffusion models face issues with monochrome inputs and high costs. DSFace, a latent diffusion-based conditional inpainting framework, addresses these challenges using a frozen Paint-by-Example diffusion model with ControlNet conditioning and DINO-V2 embeddings from a GAN-generated coarse image. Trained on the CUFS dataset, DSFace achieves superior realism, perceptual quality, and structural alignment.

1 Introduction

Generating realistic human faces from monochrome sketches is a key challenge in image-to-image translation [7, 8], benefiting domains such as character design and forensic analysis. However, the narrow distribution of single-channel sketch data and the lack of semantic cues hinder robust feature extraction and generalization. Traditional GAN-based methods [9–11] leverage semantic masks or paired supervision but often lose fine details such as wrinkles, expressions, and accessories. Diffusion models (DM) [12] and CLIP [13] have shown promise in text-to-image synthesis, yet their direct adaptation to sketch-to-image translation remains limited due to weak structural correspondence. Conditioning-guided diffusion models like ILVR [14] and SDEdit [8] use RGB references for control but still fail to preserve accurate facial geometry.

We propose a latent diffusion-based framework trained on a sketch–face dataset that reformulates sketch-to-face synthesis as a conditional inpainting problem. By leveraging the conditioning flexibility of diffusion models, we employ coarse image guidance, A PBE based baseline and DINO-V2 embeddings to provide global structural consistency. Furthermore, ControlNet and parsed facial maps enable fine-grained control over local facial attributes. The proposed model demonstrates superior realism, structural alignment, and quantitative performance on the CUFS dataset. Our contributions are as follows, 1) Reframe sketch-to-image generation as conditional inpainting using latent diffusion. 2) Employ DINO-V2 embeddings and GAN-generated coarse images for structural alignment. 3) Integrate ControlNet and semantic face parsing for fine-grained control.

*Corresponding Author.

2 Methodology

We propose solving sketch-to-face generation as a conditional diffusion inpainting task utilizing ControlNet encoder for structural control, and conditioning Dino-v2 embeddings from a coarse input generated by a GAN module.

The proposed methodology consists of two stages. In **Stage 1**, a Generative Adversarial Network (GAN) is employed to generate a coarse face image I_{coarse} from the input facial sketch I_s. The resulting coarse image I_{coarse} serves as a global structural prior, providing essential conditioning for the subsequent stage. The **Stage 2** consists of diffusion based pipeline, which consists of a frozen Diffusion Denoising Unet. The inputs to this module are face segmentation I_{seg}, face agnostic binary mask I_{mask}, face agnostic rgb image I_{ag}, input image with added noise I_t. The diffusion pipeline is guided by the structural conditioning from a ControlNet module for enabling the structural coherence with realistic facial features, the features guiding the controlNet are extracted from input sketch image I_s. An additional conditioning input that guides the ControlNet is the Dino-v2 embeddings computed from the provided coarse input image from stage 1, I_{coarse}. Dino-v2 module enables the pipeline to capture more detailed information about facial attributes, thereby enhancing control over the generation of facial regions. The ControlNet is trained to predict total noise, with the final clean generated face denoted as I_{out}.

3 Experiments

3.1 Implementation Details

We conduct end-to-end training using one NVIDIA A100 GPU with image resolution of 512×384. Training runs for 150 epochs using the AdamW optimizer

Metrics	DualGAN [1]	Pix2Pix [2]	UGATIT [3]	NICE-GAN [4]	FRAN [5]	DiSS [6]	DCNP [7]	Ours
↑SR-SIM	0.8687	0.8681	0.8758	0.8800	0.8822	0.8655	0.8787	**0.8941**
↑MSSIM	0.7938	0.7956	0.8090	0.8206	0.8381	0.8041	0.8181	**0.8618**
↓FID	126.53	85.55	117.34	84.96	73.21	86.22	65.49	**62.58**
↑FSIM	0.7893	0.7913	0.7955	0.8123	0.8187	0.7763	0.8121	**0.8352**
↑VIF	0.1128	0.1098	0.1367	0.1427	0.1798	0.1253	0.1475	**0.2321**
↑IS	1.32	1.35	1.42	1.41	1.35	1.34	1.40	**1.49**

Table 1. Comparison of various models using different metrics on CUSK datasets. Bold values indicate the best performance.

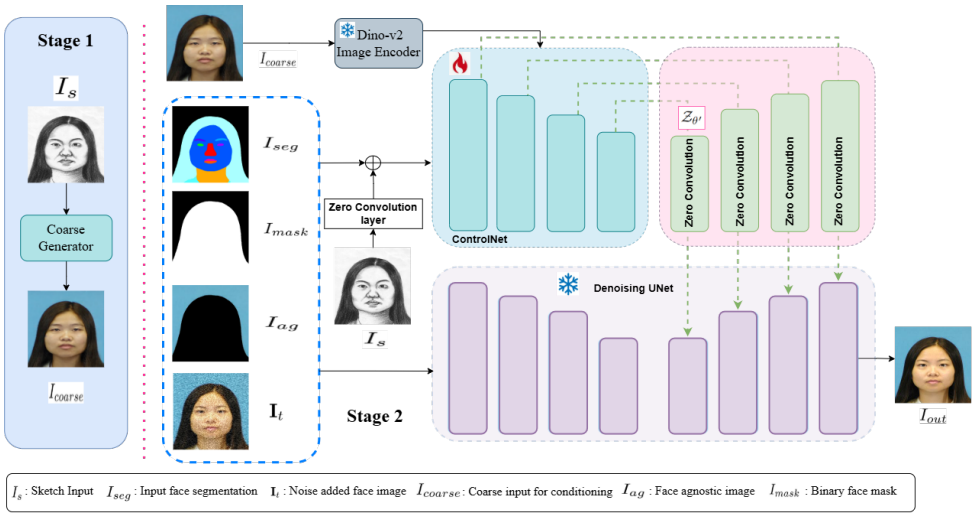

Figure 1. Architecture diagram of our proposed approach for sketch to face image synthesis

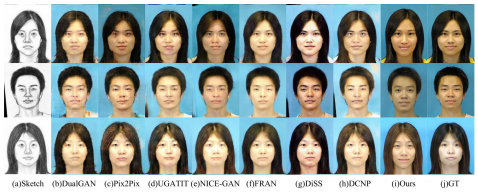

Figure 2. Qualitative comparison of our proposed approach with state-of-the-art approaches for sketch to face image synthesis.

Figure 3. Qualitative Ablation of effect of various proposed modules on quality of generated faces.

with a learning rate of 2×10^{-5}. Experiments are performed on the CUFS dataset [15], which includes 606 identities from CUHK, AR, and XM2VTS, each containing one FPS pair, split in an 80:20 ratio for training and testing.

3.2 Results

We evaluate the proposed model for sketch-to-face synthesis through qualitative and quantitative comparisons with state-of-the-art methods, including DualGAN [1], Pix2Pix [2], UGATIT [3], NICE-GAN [4], FRAN [5], DiSS [6], and DCNP [7]. As shown in Figure 2, earlier methods often yield blurry results or artifacts, while our approach produces sharper, more realistic images with well-preserved facial features and textures, benefitting from DINO-v2 and ControlNet conditioning.

Quantitatively, Table 1 highlights our approach's superior performance, achieving the highest SR-SIM (0.8941), MSSIM (0.8618), FSIM (0.8352), VIF (0.2321), and IS (1.49), along with the lowest FID (62.58). These results confirm that our approach delivers enhanced realism, structural consistency, and perceptual fidelity compared to existing approaches.

3.3 Ablation Study

We evaluate our proposed method through quantitative and qualitative ablations (Figure 3). The baseline PBE model, conditioned only on sketch input, fails to preserve structural and perceptual fidelity, yielding poor FID, MSSSIM, and IS scores. Adding a ControlNet encoder enhances facial geometry consistency, while incorporating face segmentation improves perceptual quality through semantic guidance, though texture and color remain inconsistent. Finally, conditioning ControlNet with DINO-v2 embeddings from the coarse generated image provides refined structural and appearance details.

4 Conclusion

We propose a latent diffusion model that generates realistic faces from sketches using conditional inpainting with semantic guidance. It achieves state-of-the-art realism and fidelity on the CUFS dataset, advancing sketch-based synthesis applications.

References

[1] Z. Yi, H. Zhang, P. Tan, and M. Gong. "Dualgan: Unsupervised dual learning for image-to-image translation". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2849–2857.

[2] 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.

[3] J. Kim. "U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation". In: *arXiv preprint arXiv:1907.10830* (2019).

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

[5] W. Wan, Y. Yang, S. Huang, and L. Gan. "FRAN: feature-filtered residual attention network for realistic face sketch-to-photo transformation". In: *Applied Intelligence* 53.12 (2023), pp. 15946–15956.

[6] S.-I. Cheng, Y.-J. Chen, W.-C. Chiu, H.-Y. Tseng, and H.-Y. Lee. "Adaptively-realistic image generation from stroke and sketch with diffusion model". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2023, pp. 4054–4062.

[7] M. Zhu, Z. Wu, N. Wang, H. Yang, and X. Gao. "Dual conditional normalization pyramid network for face photo-sketch synthesis". In: *IEEE Transactions on Circuits and Systems for Video Technology* 33.9 (2023), pp. 5200–5211.

[8] C. Meng, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. "Sdedit: Image synthesis and editing with stochastic differential equations". In: *arXiv preprint arXiv:2108.01073* (2021).

[9] H. Caesar, J. Uijlings, and V. Ferrari. "Cocostuff: Thing and stuff classes in context". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1209–1218.

[10] S.-Y. Chen, W. Su, L. Gao, S. Xia, and H. Fu. "Deep generation of face images from sketches". In: *arXiv preprint arXiv:2006.01047* (2020).

[11] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. "Maskgan: Towards diverse and interactive facial image manipulation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5549–5558.

[12] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[14] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon. "Ilvr: Conditioning method for denoising diffusion probabilistic models. In 2021 IEEE". In: *CVF international conference on computer vision (ICCV)*. Vol. 1. 2021, p. 2.

[15] X. Wang and X. Tang. "Face photo-sketch synthesis and recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 31.11 (2008), pp. 1955–1967.