Sketch-to-Face for more general use

Members: 陳敬涵、劉祐瑋、林士翔

Abstract

Forensic facial sketching typically relies on experienced artists to produce accurate renditions of suspects' faces—an expertise not readily available to laypersons. To address this limitation, we propose a novel sketch-to-face generation framework that transforms simple, non-professional sketches into photorealistic images. Unlike existing approaches primarily trained on Caucasian facial datasets, our method emphasizes racial diversity by incorporating a multi-ethnic dataset during training. As a result, it can robustly handle various facial attributes across different population groups. Inspired by recent advancements in generative adversarial networks, our framework employs a two-stage process: first, a sketch encoder captures essential structural features, then a conditional generator refines these features into a coherent facial image.

Introduction (Sec. 1)

In cases such as investigation or criminal cases, we now have to depend on experts. For the normals, we cannot sketch clear and similar sketches easily. Hence, we considered if we can use simple sketches that the people are able to draw to generate the corresponding pictures.

The sketch-to-face techniques now are majorly trained on Caucasian, and we would like to train on the people for more races.

Review of Previous Work (Sec. 2)

- Pix2pixHD:
    - Early works like Pix2pixHD used conditional GANs for direct sketch-to-image translation.
    - Advantages: Good results for well-aligned sketches with precise geometric correspondence
    - Disadvantages: Poor generalization to different sketch styles and freehand drawings
- DeepFaceDrawing:
    - Methods like DeepFaceDrawing decompose faces into semantic components and use manifold projection.
    - Advantages: Better handling of various sketch styles and abstraction levels
    - Disadvantages: Limited diversity in outputs and loss of fine geometric details during manifold projection
- pSp and Restyle:
    - Recent approaches like pSp and ReStyle encode sketches into StyleGAN's latent spaces (W/W+)
    - Advantages: High-quality photorealistic outputs leveraging pretrained StyleGAN

    - Disadvantages: Often fail to preserve semantic consistency with input sketches, especially for accessories like glasses and hats
- DeepFaceEditing:
    - Methods like DeepFaceEditing separate geometry and appearance control
    - Advantages: Flexible control over both structure and style
    - Disadvantages: Limited generalization to sketches different from training data
- Sketch2face
    - To improve the performance, Sketch2face uses a pre-trained StyleGAN to replace the usage of conditional GANs.
    - To solve the issue of DeepFaceDrawing, Sketch2face uses a GNN-based Sketch Semantic Interpretation for robust semantic extraction.
    - To solve the issue of pSp and Restyle, Sketch2face constructs a novel W-W+ Encoder to balance semantic control and reconstruction quality.
    - Sketch2face can't separate geometry and appearance control, but this made it can generalize to sketches which are different from training data.
-
- Summarize existing methods or research that have tackled similar problems.
- Compare and contrast any relevant advantages or disadvantages.

Technical Part

- **Sec. 3.1: Summary of the Technical Solution**
    - Paper review:

      The proposed framework consists of two main components: a Semantics-Preserving Sketch Embedding network and a Sketch Semantic Interpretation module.

        - The Semantics-Preserving Sketch Embedding network features a novel W-W+ encoder architecture with three specialized encoders:
            1. W Sketch Encoder:
                - Maps input sketches to W space for coarse semantic and geometric features.
                - Outputs a 1 × 512 vector representing high-level attributes like pose and face shape.
            2. W+ Sketch Encoder:
                - Encodes fine-level geometric and semantic details into W+ space.
                - Notably, only embeds the first 8 style codes since sketches lack appearance information.
                - The first 8 layers control geometry and semantics, while the remaining 10 layers handle appearance.
            3. W+ Appearance Encoder:
                - Generates the missing 10 style codes for fine appearance attributes.

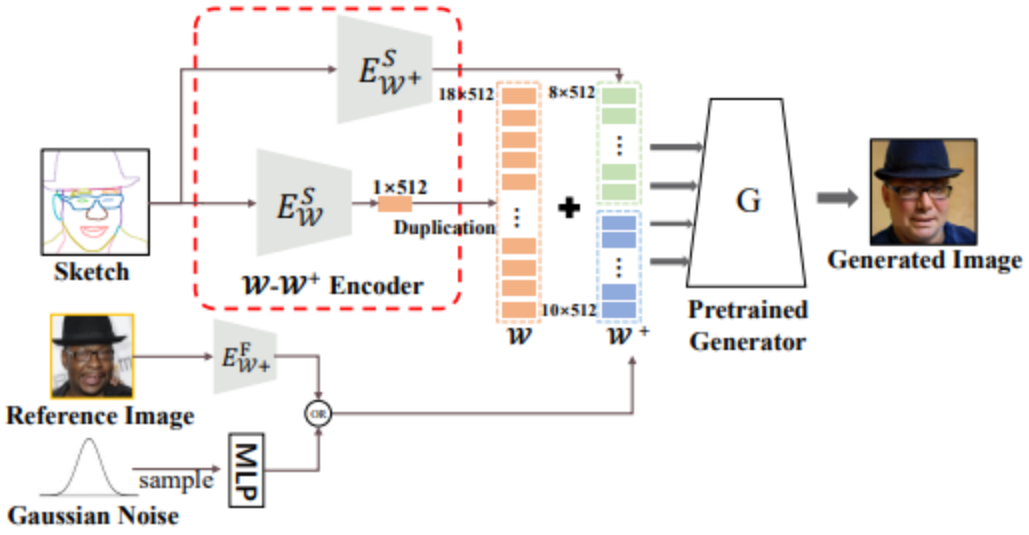
        - Can use either a reference image or random sampling for appearance control.
        - Allows flexible control over the final appearance while maintaining geometric consistency.

    The encoded features are then fed into a pretrained StyleGAN generator to synthesize photorealistic face images.

- The Sketch Semantic Interpretation module processes vectorized sketches through two key components:
    1. Stroke Structure Embedding Module (SSEM):
        - Uses three layers of bidirectional GRU.
        - Extracts structural features from individual strokes.
        - Maintains sequential ordering information within strokes.
    2. Sketch Topology Embedding Module (STEM):
        - Constructs a directed graph representing stroke relationships.
        - Uses graph neural networks to aggregate features.
        - Captures spatial topology and relationships between strokes.
        - Outputs semantic labels for each stroke.

    This two-stage architecture effectively preserves both local stroke details and global spatial relationships, enabling accurate semantic interpretation of sketches.

- Adjustment for existed method - adjust generator model
    - As noted in this paper, the W- W+ encoder is used to provide the semantic components of a draft, which are then passed to a pre-trained generator to produce realistic human faces. This model is trained on the CelebA-HQ Dataset. Subsequent experiments reveal that it demonstrates a certain robustness to human facial features, regardless of whether we provide it with an Asian face or non-human facial features. This robustness ensures that certain non-human characteristics are not transferred, which is an advantage. However, for Asian faces, it fails to effectively transfer the style to the generated images. Therefore, we decided to adjust the pre-trained generator to allow the W- W+ encoder to retain its functionality while enabling the generated images to better capture the features of the reference image, rather than being constrained by the fixed style of CelebA-HQ.



(Image above is our model architecture)

- **Sec. 3.2: Details of the Technical Solution**

  The original pre-trained generator, StyleGAN2-FFHQ-config-f, is a model trained on a human face dataset. To address its limitations, we propose several experimental methods to adjust the pre-trained generator:

  1. **Dataset Adjustment**: Train the entire model, including the W sketch encoder, on a dataset using the released checkpoint.
  2. **Utilize Alternative Pre-trained Generators**: Experiment with other pre-trained generators like StyleGAN3 and StyleGAN-XL, which offer enhanced geometric consistency, reduced aliasing, and improved detail preservation, making them better suited for handling diverse facial structures and styles.
  3. **Fine-tune StyleGAN2**: Perform fine-tuning using the original StyleGAN2 approach and StyleGAN2-ADA. The latter enhances training stability on limited or diverse datasets through adaptive data augmentation, supporting better generalization across racial and stylistic variations.
  4. **Fine-tune StyleGAN2 from Sketch**: Train on a more diverse, cross-racial facial dataset with broader styles compared to CelebA-HQ, enabling the generator to better adapt to a wider variety of reference features.

  By exploring these approaches, we aim to improve the generator's ability to preserve reference image characteristics while achieving robust and diverse style generation.

Experiments (Sec. 4)

- **Setup**:
  - Model: StyleGAN2

We opted for StyleGAN2 in our training pipeline because of its widespread acceptance in the research community and its streamlined fine-tuning process, which aligns well with our experimental objectives.

- Dataset: SCUT-FBP5500

  Because we aim to enhance the model's generalizability and current face-generation models predominantly focus on Caucasian features, we selected a dataset primarily composed of Asian individuals. Specifically, it includes 2,000 Asian males, 2,000 Asian females, 750 Caucasian males, and 750 Caucasian females.

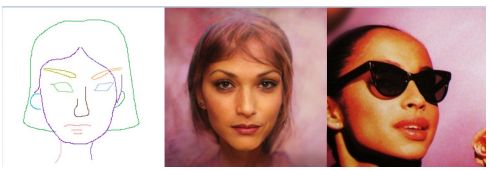
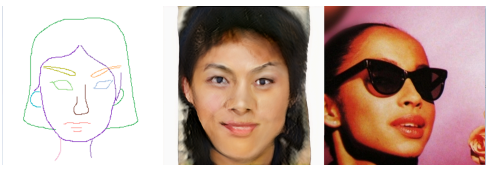
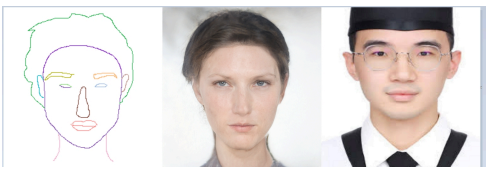
- **Results**: Present your findings in tables, plots, graphs, or images/visualizations.

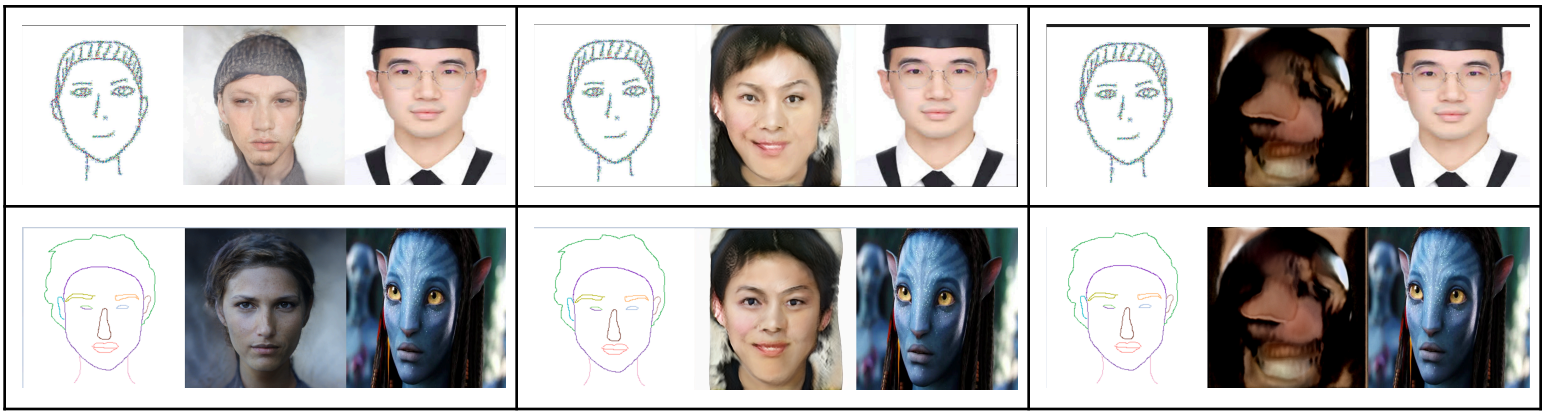
  The results of implementing the aforementioned methods are as follows:

  - For training the entire model system, it could not be executed directly due to incomplete code provided.
  - For other alternative pre-trained generators (such as other StyleGAN2 variants, StyleGAN3, StyleGAN2-ADA, and StyleGAN-XL), we conducted initial trials. However, due to differences in the original program's version and implementation, many issues arose when integrating them into the system.

  Given the limited time available for this project, we prioritized focusing on **Fine-tune StyleGAN2** and **Fine-tune StyleGAN2 from Sketch**, with the results summarized in the following table.

  (Each of the following images is a triptych. On the **left** is the sketch, on the **right** is the style reference image, and in the **center** is the synthesized result.)

Released checkpoint	Finetune stylegan2 (4000Kimg) (from checkpoint on the right)	Finetune stylegan2 (4000Kimg) (from sketch)
		
		



- **Analysis**: Discuss observations, insights, and comparisons
    - From the results shown above, we can observe that the images generated using the released checkpoint perform well for styles similar to CelebA-HQ. However, the synthesis quality is poor for Asian faces and for generating features from the sketches. On the other hand, it demonstrates strong resistance to non-human styles (using images from the movie *Avatar* as references).
    - As per the experimental design, we found that whether fine-tuning the pre-trained model or starting fine-tuning from scratch, the semantic functionality of the original model seems to be lost. It fails to generate corresponding features from the sketches, often producing nearly identical outputs.
    - Fine-tuning starting from the original pre-trained checkpoint retains facial features and allows for a broader range of styles, but the overall performance still needs improvement.
    - **Insight:** In the future, it would be beneficial to train the entire system to avoid losing some of the original functionalities. Additionally, during fine-tuning, retaining CelebA-HQ in the training process could help prevent style deviations that lead to distortions.

Conclusions (Sec. 5)

From this implementation, we gained valuable insights into the state-of-the-art (SOTA) process for sketch-to-face generation. We also explored its application in more practical scenarios and identified its limitations. Through this fine-tuning experiment, we observed that while the paper suggests that the pre-trained model achieves disentanglement, in practice, the disentanglement is not always clean or robust. This highlights the necessity of training the entire system to achieve optimal performance and functionality.

This project allowed us to make incremental progress and overcome some challenges, shedding light on the potential for improvement. Looking ahead, we are excited about future opportunities for integrating and applying such models across diverse datasets. We hope these efforts will lead to deeper insights and broader breakthroughs, especially in expanding the

versatility of sketch-to-face generation systems to handle various styles, domains, and real-world applications.

References

Yang, B., Chen, X., Wang, C., Zhang, C., Chen, Z., & Sun, X., "Semantics-Preserving Sketch Embedding for Face Generation." IEEE Transactions on Multimedia, 25, 8657-8671 (2022).

] P. Zhu, R. Abdal, Y. Qin, J. Femiani, and P. Wonka, “Improved stylegan embedding: Where are the good latents?” arXiv preprint arXiv:2012.09036, 2020.

Stylegan series
https://github.com/NVlabs/stylegan2
https://github.com/NVlabs/stylegan2-ada-pytorch
https://github.com/NVlabs/stylegan3
https://github.com/autonomousvision/stylegan-xl