FaceSketching: Interactive Realistic Face Image Creation from Free-hand Sketches

Anonymous Author(s)
Submission Id: 1234

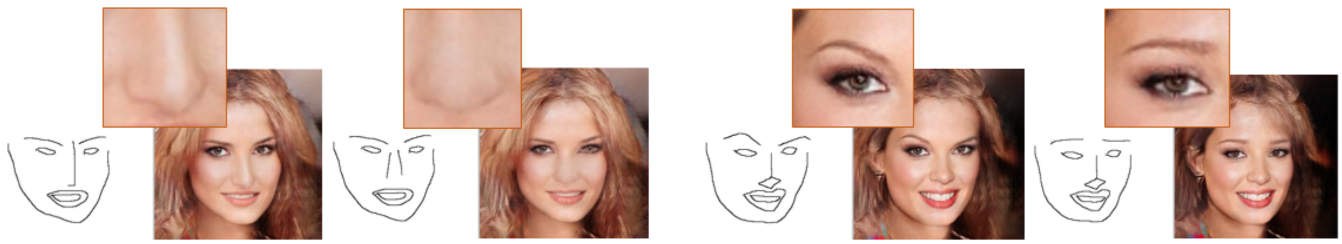

Figure 1: This is a teaser

ABSTRACT

In this paper, we explore the task of generating photo-realistic face images from sketches based on image-to-image translation framework. Since there exists no large-scale dataset of face sketches, existing methods utilize edge maps of face images as training data. However, edge maps perfectly align with edges of the corresponding face images, which limit existing models' generalization to hand-drawn sketches with vast stroke diversity. To address this problem, we propose a robust sketch-to-face translation model which is able to generate photo-realistic face images from hand-drawn sketches. A novel module, named spatial attention pooling (SAP) is designed to adaptively handle stroke diversity. We conduct extensive experiments on CelebA-HQ dataset. The experiment results show the superiority of our model over existing methods on perceptual realism and generalization.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**.

KEYWORDS

Image synthesis, sketch-based interface, face editing, deep neural network

ACM Reference Format:
Anonymous Author(s). 2020. FaceSketching: Interactive Realistic Face Image Creation from Free-hand Sketches . In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Flexibly creating new content is one of the most important goals in both computer graphics and computer-human interaction. While sketching is an efficient and natural way for common users to express their ideas for designing and editing new content, sketch-based interaction techniques have been extensively studied since the very early stage of computer graphics [**? ? ? ? ?**]. Imagery content is the most ubiquitous media with a large variety of display devices everywhere in our daily life. Creating new imagery content is one way to show people's creativity and communicate smart ideas. In this paper, we target portrait imagery, which is inextricably bound to our life, and present a sketch-based system that allows common users to create new face imagery by specifying the desired facial shapes via free-hand sketches (sketch-to-face).

Deep learning techniques bring significant improvement on the realism of virtual images. Recently, a large number of studies have been conducted on general image-to-image translation which aims to translate an image in one domain to a corresponding image in another domain, preserving the same the content, such as structure, scene or objects [3, 5, 7, 9**?** −11]. Treating sketches as the source image domain and realistic face images as the target domain, image-to-image translation techniques can be directly applied to our task.

Since there exists no large-scale face sketch dataset and collecting hand-drawn sketches is time-consuming, existing methods [3, 6, 7] applied on sketch-to-face utilize edge maps or contours of real face images as training data. Edge maps and contours enable existing models to be trained in a supervised manner and obtain plausible results on their test sets. However, models trained on edge maps and contours are not able to achieve satisfactory results on hand-drawn sketches, specially on those drawn by users without drawing skills.

Since strokes in edge maps and contours align perfectly with edges of the corresponding real images, models trained on edge-aligned data tend to generate unreal shapes of facial parts following the inaccurate strokes when the input sketch is poor-drawn. Hence, for an imperfect hand-drawn sketch, it is a trade-off between *the realism* of the synthesized image and *the correspondence* between input sketch and the edges of the synthesized image. Models with high edge-alignment fails to be generalized to sketches with imperfect strokes.

Moreover, we observe that the balance between the trade-off mentioned above varies from one position to another across the image. In a sketch, some facial parts might be well-drawn while the others not. For the well-drawn facial parts, the balance are supposed to move towards the correspondence ensuring those parts in synthesized image depicting the user's imagination. On the other hand, the areas of poor-drawn parts should emphasize the realism and not follow the irregular shapes and strokes.

Based on the discussion above, we propose a novel sketch-based synthesis framework which is robust to hand-drawn sketches. A new module, named spatial attention pooling (SAP), is designed to adaptively adjust the spacial-variant balance between *realism* and *correspondence* across the image. In order to break the edge-alignment between sketches and real images, SAP relaxes strokes with one-pixel widths to multiple-pixel widths using pooling operators. A larger width of a stroke, which is controlled by the kernel size of pooling operator, indicates the less restrict between this stroke and the corresponding edge in the synthesized image. However, the kernel size is not trainable using back propagation algorithm. Hence, for an input sketch, four branches of pooling operators with different kernel sizes are added in SAP to get four relaxed sketches with different widths. The relaxed sketches are then fused by a spatial attention layer which adjusts the balance of *realism* and *correspondence* . For each position, the spatial attention layer assigns high attention to the relaxed sketch with large width if this position requires more *realism* than *correspondence*.

In summary, our contribution in this paper is three-fold.

- Based on comprehensive analysis on the edge alignment issue in image translation frameworks, we propose a sketch-to-face translation system that is robust to hand-drawn sketches.
- A novel deep neural network module for sketch, named spatial attention pooling, is designed to adaptively adjust the spatial-variant balance between the realism of the synthesized image and the correspondence between the input sketch and the synthesized image.
- Extensive experiments demonstrate the superiority of our model over existing methods on perceptual realism and generalization.

2 RELATED WORK

Our method is related to studies on image-to-image translation, sketch-based image generation and face image generation and editing. In this section, we discuss the most related works of our method.

2.1 Image-to-Image Translation

Given an input image from one domain, an image-to-image translation model outputs a corresponding image from another domain and preserves the content in the input image. Existing image-to-image translation models are based on generative adversarial networks conditioned on images. Pix2pix [3] is the first general image-to-image translation model which is able to be applied to different scenarios according to the paired training images, such as, semantic maps to real images, day images to night images, image coloring, and edge maps to real images. [4] utilizes semantic label maps and attributes of outdoor scenes as input and generates the corresponding photo-realistic images. In order to model multi-modal distribution of output images, BicycleGAN [**?**] encourages the connection between the output and the latent code to be invertible. CycleGAN [**?**], DualGAN [**?**], and DiscoGAN [**?**] propose unsupervised image translation model with a similar idea named cycle consistency borrowed from language translation literature. StarGAN [**?**] presents a framework for one-to-many image translation by adding a domain code as input and a domain classifier as guidance. Pix2pixHD [7] is proposed as a high-resolution image-to-image translation model for generating photo-realistic image from semantic label maps using a coarse-to-fine generator and a multi-scale discriminator. It can also be applied to edge-to-photo generation by using the paired edge maps and photos as training data. However, xxxxxxxxxxxxxx

2.2 Sketch-based Image generation

Sketch-based image generation is a hot topic in computer vision and computer graphics. Given a sketch of a scene with text labels for objects, traditional methods, such as Sketch2Photo [1] and PhotoSketcher [2], search corresponding image patches from a large image dataset and then fuse the the retrieved image patches together according to the sketch. These methods are not able to ensure the global consistency of the resultant image and fails to generate totally new images. After the breakthrough made by deep neural networks (DNN) in computer graphics and computer vision, a variety of DNN-based models have been proposed for sketch-based image generation. The general image-to-image translation models mentioned above are able to be applied to sketch-based image generation once sketches and the corresponding images are used as training data. Besides, many other models are designed specially for sketch inputs. SketchyGAN [**?**] aims to generate real images from multi-class sketches. A novel neural network module, called mask residual unit (MRU), is proposed to improve the information flow by injecting the input image at multiple scales. Edge maps are extracted from real images and utiled as training sketches. However, the resultant images of SketchyGAN are still not satisfied.

2.3 pooling

2.4 Face Generation and Editing

All existing methods for local face editing require the user to provide masks and strokes manually. In comparison, strokes are only the input to indicate the desired shape. Our system automatically interprets the intended edit and produces the local change accurately. This greatly reduces users' burden and preserves the fluency of user interaction.

Only local edit in a relatively small area is supported. As reported in SC-FEGAN [], artifacts appear when complete a large region in FaceShop [].

3 DEEP NETWORK FOR SKETCH-PHOTO TRANSLATION

Sketches drawn by users without well-trained drawing skills always contain severe line distortion, which may bring large degeneration to the generated face images and therefore damage the their quality hugely. For this reason, we propose a sketch-to-photo translation model that is able to adaptively handle the line distortion of input sketches.

The architecture of the proposed model is shown in Figure 2. We use pix2pixHD model [7] with the 'global' version generator and the multi-scale discriminator as baseline model. The input of the generator is either a face sketch or a deformed face sketch and the output is the corresponding generated photo-realistic face image. A novel module named Spatial Attention Pooling (SAP) is added in the front of the generator to enable our model to adaptively handle line distortions of input sketches (before the lowest level of feature extraction.) For the discriminator, the generated image concatenated with the input sketch (or the deformed input sketch) is treated as a fake sample while a real face image sampled from real face distribution concatenated with its corresponding sketch is regarded as a real sample. In order to guided the model with SAP to be tolerant with line distortion of input sketches, we design a novel generator feature matching loss for our task, besides the adversarial loss, the reconstruction loss and the discriminator feature matching loss used in pix2pixHD.

In this section, we first introduce the dataset we use in our experiment in Subsection **??**. Then we describe the proposed SAP in Subsection 3.0.2. At last we discuss losses we apply in Subsection 3.1.1.

3.0.1 Face Sketches and Stroke Deformation.

3.0.1 Face Sketches and Stroke Deformation. Paired face sketch-photo dataset is required for supervised sketch-to-face translation methods. Since there exits no large-scale paired face sketch dataset, the training face sketches used by existing methods are generated from face image dataset, e.g. CelebA-HQ face dataset, using edge detection algorithm such as HED [8]. However, the sketches generated by edge detection algorithm are sometimes incomplete or other problems. Discuss the advance of make-edges over edge maps and contours (I would say the edge maps are quite different from handdrawn sketches, not because of the incompleteness.) Although CSAGAN [?] applied self-attention mechanism to alleviate the incompleteness problem, the others remains. Therefore we use another method to generate clearer and complete sketches from face images.

[7] introduce another method to generate sketches from face images. Given a face image, the face landmarks are detected using an off-shelf landmark detection model. A new kind of sketch, denoted as *face contour*, is obtained by connect specific landmarks. Sketch-to-face model trained by face contour fail to generalize to hand-drawn sketches with hair, wrinkles or beard.

The CelebAMask-HQ dataset [?] provides a face semantic map for each face image in CelebA-HQ dataset. We basically use the boundary map of the semantic map as the sketch of the corresponding face image. Figure 3 shows an example of comparison between a sketch generated by edge detector, a face contour and a sketch generated from semantic boundary.

Stroke Deformation. A shortcut of sketches generated from semantic boundary (and those generated by edge detector) is the lines of sketches are perfectly aligned to edges of the corresponding face images. In order to break the edge-alignment and mimic the hand-drawn sketches, we apply a deformation to the lines, similar to that in FaceShop [?]. Specifically, we vectorize lines of each sketches using AutoTrace algorithm [?], and add an offset randomly selected from $[-d, d]^2$ to the control points and end points of the vectorized lines, where d is the maximum offset and we set $d = 11$ in our experiments. Both the initial sketches generated from semantic boundary and deformed sketches are used as the input sketches to our model.

3.0.2 Spatial Attention Pooling. (Check if the idea about the relax is well-described) (Spatial-specific?Position-specific?Spatially variant?) When a input hand-drawn sketch is not well-drawn, line distortions of the sketch damages the quality of the generated face image. It is a trade-off between the realism of the output face image and the alignment between the input sketch and the output face image. In order to alleviate the edge alignment between the input sketch and the output face image, we desire to relax thin strokes to ambiguity bands with various width or uncertainty. One of the straightforward ways is to smooth the lines of sketches using image smoothing algorithm. Another is to dilate the sketch lines so that the widths of lines are of multiple pixels [?]. However, the capacity of either the two hand-crafted ways above is limited, because the uniform smoothness and the dilate radius for all positions of the whole sketch violate the unevenness of hand-drawn sketches on depicting different facial parts. We argue that the balance between the realism of the output face image and the alignment between the input sketch and the output face image differs from one position to another across the face image. Therefore, the relax degree should be spatial-specific.

Based on the discussion above, we propose a new module, called spatial attention pooling (SAP), to adaptively relax the widths of strokes in the input sketch in a spatial-specific way. (I would not use 'dilate' since this simple word does not reveal the underlying discovery.) Given an input sketch $S \in \mathbb{R}^{H \times W}$, we first pass it through N_r pooling branches with different kernel sizes of $\{r_i, i = 1, \dots, N_r\}$ to get $\{P_i | i = 1, \dots, N_r\}$. In order to get a relaxed sketch $\tilde{S}$ with spatial-specific stroke widths, we compute

$$\tilde{S} = \sum_{i=1}^{N_r} A_i * P_{r_i}(s), \tag{1}$$

where $*$ is element-wise multiplication and A_i is the ith channel of the spatial attention map $A \in \mathbb{R}^{N_r \times H \times W}$. The spatial attention map A controls the relax degrees of all positions of the input sketch. Since a line with a large distortion is supposed to be assigned with a large relax degree, A is supposed to adaptively pay more attention (a large value) to a P_i with a large kernel size in the areas with large line distortions.

A straightforward way to get A is passing the input sketch through a few convolutional layers and these convolutional layers are trained to detect the areas with line distortions. However, we found the a few convolutional layers are insufficient to learn to detect line distortions directly. Therefore, we introduce a two-class classifier to ease the detection. Specifically, we pre-train a fully-convolutional

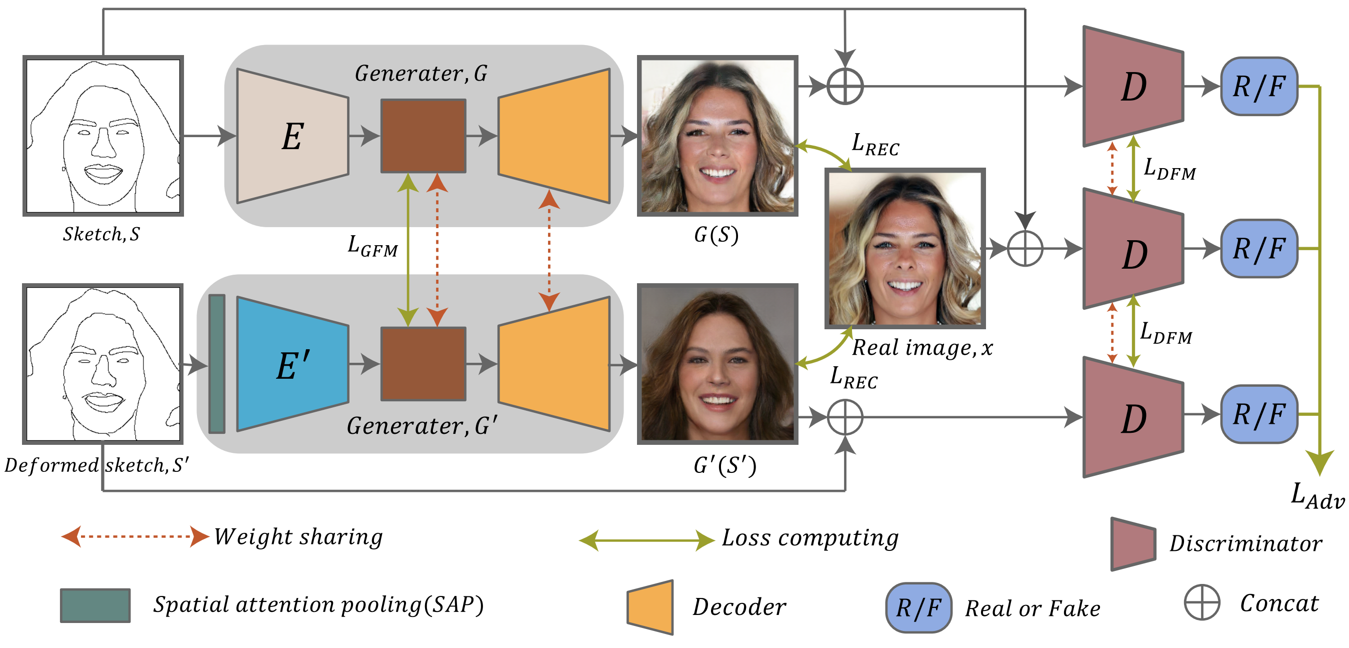

Figure 2: The architecture of our model.

Figure 3: Comparison between a sketch generated from edge detection and from semantic boundary.

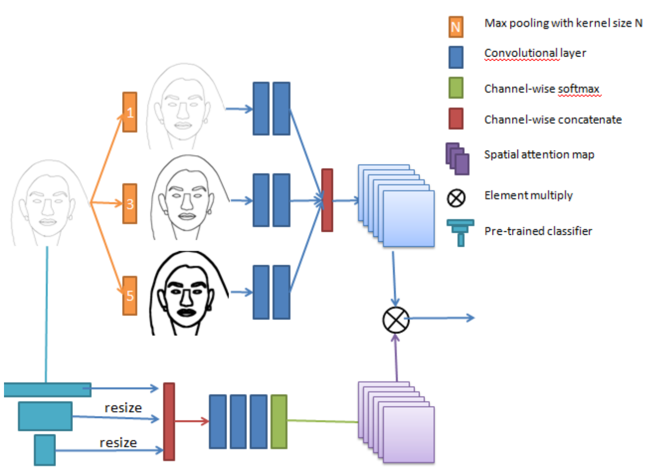

Figure 4: Sap

two-class classifier C with three convolutional layers to distinguish sketches from deformed sketches. Then we utilize this pre-trained classifier to extract features of the input sketch S to get $C_i(S), i = 1, 2, 3,$ where $C_i()$ denotes the ith feature maps extracted by C. These feature maps from classifier emphasize the differences between sketches and deform sketches. We resize and concatenate these feature maps, and pass them through three convolutional layers to get the spatial attention map:

$$A = Softmax(Conv([C_1, Up_2(C_2), Up_4(C_3)])), \tag{2}$$

where Up_2 and Up_4 indicates $2\times$ and $4\times$ upsampling, Conv($\cdot$) indicates three cascaded convoutional layers, and $Softmax()$ is a softmax layer computed over channels to ensuring that for each position of A, the sum of weights of all channels equals to 1.

We show this idea in Fig. 4.

3.1 Traning Schedule

3.1.1 Losses. generator feature matching effect and losses summary

(Before describe how you do it, please describe why you do this.)

Let $G_q(\cdot)$ produces the feature maps of the q-th layer in the generator G. Given an input sketch S and the corresponding deformed sketch $\tilde{S}$, we compute the generator feature matching loss as:

$$\mathcal{L}_{GFM}(G) = \mathbb{E}_{S\sim p_{data}(S)} \frac{1}{N_Q} \sum_{q\in Q} \frac{1}{|G_q|} \|G_q(S) - G_q(\tilde{S})\|_1, \tag{3}$$

where $|G_q|$ denotes the number of elements in $G_q(\cdot)$, Q indicates a set of the selected generator layers for computing this loss and the size of Q is N_Q. We select the xxx layers of the generator in our experiments.

Besides the generator feature matching loss $\mathcal{L}_{GFM}(G)$, for generator G and multi-scale discriminator $D = D_k|k = 1, 2, ..., N_D$, the adversarial loss $\mathcal{L}_{GAN}(G, D)$ and the discriminator feature matching loss $\mathcal{L}_{DFM}(G, D)$ are computed as the same form as those in

Figure 5: Sketch interface. Full interface with editing tools.

pix2pixHD [7]. Discussion: add equations of these two losses or not The objective of the proposed model is:

$$\min_G \max_D \mathcal{L}_{GAN}(G,D) + \lambda \mathcal{L}_{DFM}(G,D) + \mu \mathcal{L}_{GFM}(G). \quad (4)$$

where λ and μ are the weights for balancing different losses. We set $\lambda = xxx$ and $\mu = xxx$ in our experiments.

4 EXPERIMENTS

4.1 Sketch Interface

We develop a sketch-based interface which allows common users to describe their desired face and part shapes by a few strokes. Once the user finishes his drawing, the generated face image is shown on the right bottom of the sketch? This will be helpful for local editing., as Fig. 5 shows. The user is allowed to edit the sketch by erasing strokes or adding new stroke to change eye shapes, noses shapes, eyebrows, and so on. Each round of face image generation after users' modification takes about ... seconds on a XXX GPU with xxGB memory.

4.2 Face Generation from Contours

Contour Dataset. (Yangbinxin: Add description on the dataset generation.)

Photo Generation from Contours. The Pix2pixHD [] was applied to the task of translating edges to photos. They successfully generate hiqh-quality photos from contours that are generated from real photos. However, when we apply the Pix2pixHD model with hand-drawn sketches which present distinct characteristics from the synthesized contours which are smooth and clean without geometric distortions, it fails to produce good results, as Fig. 6 shows. (Add more results with other models.) We can see that (expected results: (e) have better details. (d) i have no idea. c is better than (b).)

Face Editing with Strokes. When users want to change local shapes of facial features, a few strokes can be modified in our interface. The pix2pixHD model does not take the sparsity of sketches and the instance normalization tends to normalize local regions with a global style factor, which mainly conveys brightness variance. Therefore, when local details such as hair, the bamoustache, and so on, the results change significantly even only local region is modified. In comparison, our proposed SLRN captures the shape details in the drawn sketches and successfully avoid edge-aligned artifacts caused by distortion in hand-drawn sketches, as Fig. 7 shows.

Figure 6: Face generation with different models. From left to right: (a) hand-drawn sketches as input. (b) Results generated by pix2pixHD that is retrained using our contour-photo dataset. (c) Results generated by pix2pixHD that is retrained using our contour-photo dataset with geometric transformation as data augmentation. (d) Results generated by removing instance normalization at the shallow convolution layers (five layers in the global generator). (e) Results from our model (pix2pixHd architecture by replacing instance normalization with the proposed SLRN.) More results can be found here: (Provide a link for all results.)

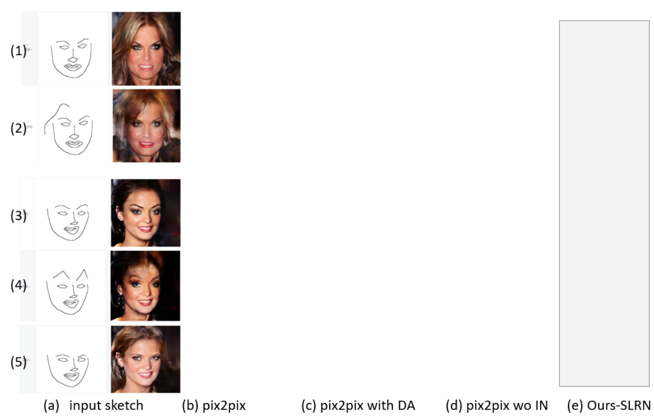

Figure 7: Local face editing with different models. The proposed SLRN captures the shape details in the drawn sketches and successfully avoid edge-aligned artifacts caused by distortion in hand-drawn sketches. More results can be found here: (Provide a link for all results.)

In order to analyze the effect of instance normalization and our SLRN, we visualize the features extracted at early stages in the generators of different models. As Fig. 8 shows, ...

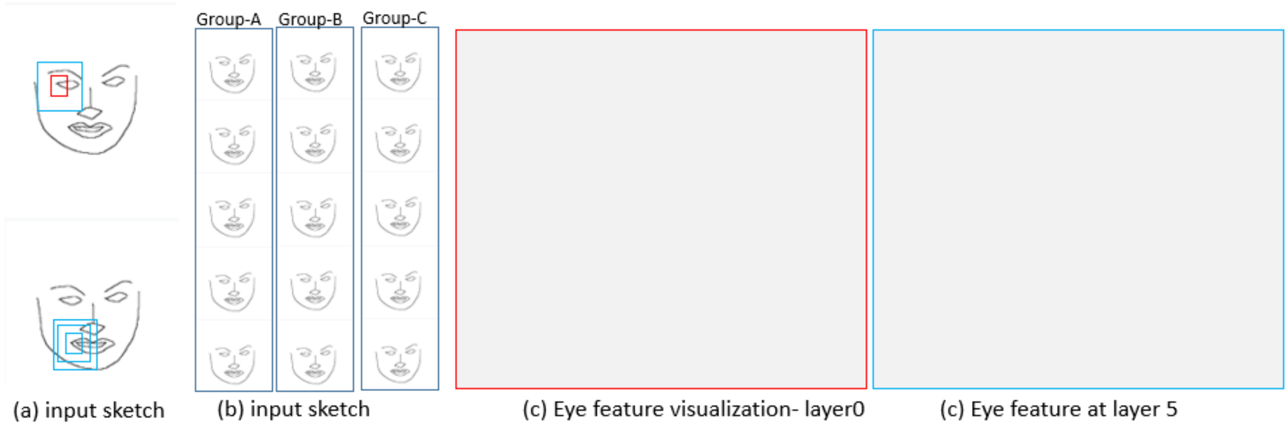

Figure 8: Visualization of extracted features under local editing. We extract the features at different convolution layers at the left eye position (a) from three groups of sketches (b) with different types of local editing. The extracted high-dimensional features using different models including pix2pixHD-DA, pix2pixHD-wo-IN, and our SLRN) are mapped into 2D space using TSNE [?] in (c). (ChengZhiHua: provide a link for all results.)

4.3 Comparison with Image Translation networks

Existing DNNs for image translation can be trained for sketch-photo translation using the paired dataset.

4.4 Comparison with Image Editing

4.5 Limitations and future work

REFERENCES

[1] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. 2009. Sketch2Photo: internet image montage. *ACM Trans. Graph.* 28, 5 (2009), 124:1–124:10.

[2] M. Eitz, R. Richter, K. Hildebrand, T. Boubekeur, and M. Alexa. 2011. Photosketcher: Interactive Sketch-Based Image Synthesis. *IEEE Computer Graphics and Applications* 31, 6 (Nov 2011), 56–66.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5967–5976. https://doi.org/10.1109/CVPR.2017.632

[4] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. 2016. Learning to Generate Images of Outdoor Scenes from Attributes and Semantic Layouts. *CoRR* abs/1612.00215 (2016). arXiv:1612.00215

[5] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192* (2017).

[6] Yuhang Li, Xuejin Chen, Feng Wu, and Zheng-Jun Zha. 2019. LinesToFacePhoto: Face Photo Generation From Lines With Conditional Self-Attention Generative Adversarial Networks. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. ACM, New York, NY, USA, 2323–2331. https://doi.org/10.1145/3343031.3350854

[7] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8798–8807. https://doi.org/10.1109/CVPR.2018.00917

[8] Saining Xie and Zhuowen Tu. 2015. Holistically-Nested Edge Detection. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1395–1403. https://doi.org/10.1109/ICCV.2015.164

[9] Z. Yi, H. Zhang, P. Tan, and M. Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2868–2876. https://doi.org/10.1109/ICCV.2017.310

[10] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2242–2251. https://doi.org/10.1109/ICCV.2017.244

[11] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward Multimodal Image-to-Image Translation. In *Advances in Neural Information Processing Systems 30*. 465–476.

REFERENCES

[1] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. 2009. Sketch2Photo: internet image montage. *ACM Trans. Graph.* 28, 5 (2009), 124:1–124:10.

[2] M. Eitz, R. Richter, K. Hildebrand, T. Boubekeur, and M. Alexa. 2011. Photosketcher: Interactive Sketch-Based Image Synthesis. *IEEE Computer Graphics and Applications* 31, 6 (Nov 2011), 56–66.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5967–5976. https://doi.org/10.1109/CVPR.2017.632

[4] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. 2016. Learning to Generate Images of Outdoor Scenes from Attributes and Semantic Layouts. *CoRR* abs/1612.00215 (2016). arXiv:1612.00215

[5] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192* (2017).

[6] Yuhang Li, Xuejin Chen, Feng Wu, and Zheng-Jun Zha. 2019. LinesToFacePhoto: Face Photo Generation From Lines With Conditional Self-Attention Generative Adversarial Networks. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. ACM, New York, NY, USA, 2323–2331. https://doi.org/10.1145/3343031.3350854

[7] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8798–8807. https://doi.org/10.1109/CVPR.2018.00917

[8] Saining Xie and Zhuowen Tu. 2015. Holistically-Nested Edge Detection. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1395–1403. https://doi.org/10.1109/ICCV.2015.164

[9] Z. Yi, H. Zhang, P. Tan, and M. Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2868–2876. https://doi.org/10.1109/ICCV.2017.310

[10] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2242–2251. https://doi.org/10.1109/ICCV.2017.244

[11] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward Multimodal Image-to-Image Translation. In *Advances in Neural Information Processing Systems 30*. 465–476.