Rebuttal

We sincerely thank the anonymous reviewers for their positive feedbacks and insightful comments. We now address the concerns point by point in detail.

[Review1-Q1]: The experimental results section could do with more details on the number of hand-drawn sketches.

[A1.1]: We invited \*\*one\*\* expert with well-trained drawing skills to draw \*\*12\*\* portrait sketches for testing. Also, we invited \*\*20\*\* graduate students without drawing skills to \*\*200\*\* draw freehand sketches. We would like to share all our data and code in the future.

[Review1-Q2]: Are results in Figure 6/7 representative of the entire set, or only the best results?

[A1.2]: Basically, most of our results on both types of hand-drawn sketches surpass those of both baseline methods in generating fine textures and realistic face shapes. We selected results shown in Figure 6/7 to better demonstrate our model’s effectiveness on spatially adjusting balance between realism and conformance, which is one of the key ideas of the proposed method. See [A3.3] for further explanation on this idea. We will emphasize this idea in the descriptions of Figure 6/7.

[Review2-Q1]: Comments on typos and minor errors.

[A2.1]: Thank you for your detailed comments. We will correct all typos and errors..

[Review3-Q1]: Can the proposed solution generalized to other types like animals, cars, etc. how and why?

[A3.1]: We currently focus on front images, which are aligned and show the unified structure. The proposed solution might be easy to be generalized to the datasets with aligned images of other types, such as animal face dataset [MUNIT, Huang et al. ECCV 2018], shoes dataset and handbag dataset [pix2pix, Phillip et al. CVPR 2017]. However there potentially exists some challenges in generalizing to unaligned dataset or dataset with multiple object classes, like ImageNet, due to the diversity of structures, shapes and scales.

[Review3-Q2]: Comparison with paper Chen et al. arXiv 2020.

[A3.2]: We did not compare our method with Chen et al 2020 in our submission since it was first available at arXiv in June 2020, AFTER the submission date of our paper. Both our method and the method proposed in Chen et al. arXiv 2020, denoted as DeepFaceDrawing, train generators with synthesis sketches and test with hand-drawn sketches. DeepFaceDrawing separates face parts into patches to add local-global soft constraint on local parts, while our method uses multiple branches of pooling layers and spatial attention mechanism to achieve spatially varying balance adjusting. Besides, DeepFaceDrawing utilizes manifold projection to align feature maps between hand-drawn sketches and synthesis sketches, while our method uses dual generator architecture and generator feature matching loss to achieve the same goal.

[Review3-Q3]: 'We argue that the balance between the realism and the conformance differs from one position to another across the face image'. This is not clear. Can you explain why?

[A3.3]: For most generators trained in CGAN framework and reconstruction losses, generating the results with strict edge alignment leads to lower reconstruction loss. However, if the input freehand sketches have distortions, strict edge alignment leads to low realism of the generated images. Even in the same input sketch, the user might draw strokes in some parts (such as, eyes, mouth) carefully with little distortions, while draw strokes in other parts (chin, hair) roughly. Therefore, a good generator should be able to adaptively adjust the balance between the edge conformance and the realism of the generated images for different parts.

The balance moves forwards to the conformance at a well-drawn part with realistic shape to meet the desire/tendency of user, while moving forwards to the realism at a poorly-drawn part to ensure the quality of the generated image.

[Review3-Q4]: Eq.8 is not clear.

[A3.4]: Thanks for pointing this out. We will carefully proofread our manuscript and make all symbols and equations consistent.

[Review3-Q5]: The proposed solution is very expensive in term of complexity because it requires treating each facial feature locally first and then the face as a whole? can you elaborate more on this?

[A3.5]: The complexity is not as large as it appears since the number of channels of feature maps in SAP is small (48). In the test stage, we only use the generator with SAP. Compared to the generator without SAP (parameters 102.6M, FLOPs 30.9G), our SAP-Generator including the pretrained classifier has parameters 103.2M, FLOPs 42.6G. We will add this comparison in our paper.

[Review3-Q6]: More datasets and comparison are needed for the evaluation.

[A3.6]: More experiments and comparisons on other datasets are to be added in revised paper.

以上字符数：4687 （Rebuttal要求最大字符数是5000）

草图的具体数量还要根据最后的收集的数量修改。