Supplementary Materials for Aggregation via Separation: Boosting Facial Landmark Detector with Semi-Supervised Style Translation

Paper ID 1870

The content of our supplementary material is organized as follows.

1. More ablation studies and detailed analysis of components in our framework.
2. Additional discussion about related directions.
3. Details of our annotated AFLW-68 dataset and some representative visualized samples.

S1. More Ablation Studies

In this section, we provide additional analysis about each design in our framework to facilitate understanding of our structure. Two key loss terms in our framework are studied to give insights into their respective roles. Qualitative visualization and quantitative results are reported for a comprehensive comparison.

KL divergence loss and perceptual loss, are incorporated into our framework during the disentangled learning procedure. Fig. 1 shows their respective effect on style translation via visual comparisons of several incomplete variants. Through visual observations, their roles could be inferred intuitively. The perceptual loss, as discussed, is designed to capture better style information and visual quality. Thus, removing this term leads to "over-smoothness" and poor diversity on synthetic images. Removing KL divergence term shows severe structure distortion on translated results, which indicates that KL divergence loss plays a key role on disentangling structure and style information.

Quantitative results of each variants are also reported in Table. 1. The normalized mean error(NME) is evaluated on WFLW [11] test set when the model is trained on style augmented dataset using each variant. We observe that NME will increase if any loss function is removed. In particular, the detector performance drops significantly lower than the baseline if L_{KL} is removed. Both the qualitative and quantitative result interprets the role of each component, indicating their essentialness in our framework.

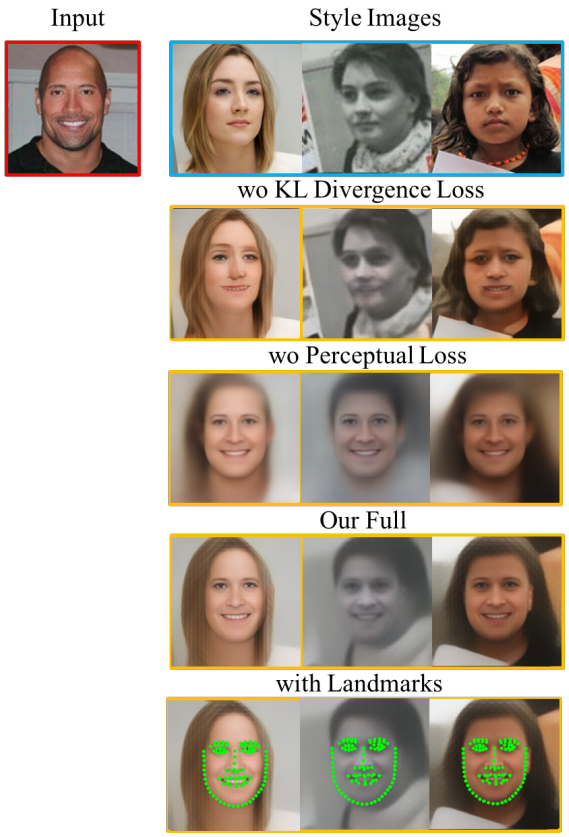

Figure 1: Qualitative analysis of each component in our framework. Given input images in red, 3 different styles are provided to perform translation towards input structure. 3 incomplete variants of our framework are used to show the functionality of each component.

Model	Baseline	wo KL divergence	wo Perceptual	Full
NME(%)	8.49	9.08	8.34	**7.98**

Table 1: Quantitative ablative results. Normalized mean error (%) on WFLW test set using different variants of our framework.

S2. Additional Discussion

In this section, we provide more discussion on our approach along with our analysis towards some existing alternatives.

S2.1 Comparison with GAN-based approaches

Generative adversarial network (GAN) and its applications are widely studied these days, using GAN-synthetic

data to aid training, has also been explored along this line. Some works [1] have utilized GANs to perform data augmentation. However, its effect still remains questionable especially on high-level vision challenges. For instance, in our task, face images need to be labeled with accurate landmarks. Existing generative models are incapable of handling these tasks with fine-grained annotations, e.g. semantic segmentation, constrained by its limited generalizability. We choose to escape the difficulties of GAN training, starting from a new perspective of internal representation. With decent representation of separating style and structure, different interactions within a face image can be simulated by re-rendering from existing style and structure code. In other words, our choice depends upon fully exploiting available information by mixing them, instead of creating new information and visually perfect results via adversarial learning procedure. However, if two codes of structure and style are factored well, advances on high fidelity images synthesis could theoretically bring more gains based on our framework.

S2.2 Comparison with Style Transfer

Our method is motivated by advances in style transfer. A common doubt could be why not directly conducting style transfer as a augmentation or how basic style transfer could help training. As discussed, our definition of style includes environments and degradation that prevent the model from recognizing while content refers to facial geometry. Applying "vanilla" style transfer would leads to structural distortion on stylized images, as illustrated in Fig. 2.Our definition of "style" helps preserve structure on synthetic images. Besides, synthetic images using style transfer have a large domain gap with real-world face images. Simply augmenting training with these samples would instead hurt model's localization ability on real images.

Figure 2: Visual comparison with style transfer approach. For the style transfer algorithm, we use [3]. Our results are more realistic than stylized images, with better structure coherence.

S3. Details of AFLW 68-point dataset

We propose a new facial landmark dataset based on AFLW [6], to facilitate benchmarking on large pose performance. To allow a more precise evaluation and cross-dataset comparison, we follow the widely-used Multi-PIE [5] and 300W [10] 68-point protocol. Annotated samples are provided at Fig. 3, which contains extreme pose variations.

Database	Environment	Number
Multi-PIE [5]	Controlled	750000
XM2VTS [8]		2360
LFPW [2]		1035
HELEN [7]	In-the-wild	2330
AFW [9]		468
IBUG		135
COFW-68 [4]		507
AFLW-68(Ours)		**25993**

Table 2: Widely-used 68-pt facial landmark datasets. Dataset names their the environment and number are reported.

References

[1] A. Antoniou, A. Storkey, and H. Edwards. Data augmentation generative adversarial networks, 2018. 2

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013. 2

[3] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 2

[4] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *CVPR*, 2014. 2

[5] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 2

[6] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV workshops*. IEEE, 2011. 2

[7] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, 2012. 2

[8] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, 1999. 2

[9] D. Ramanan and X. Zhu. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 2

[10] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV Workshops*, 2013. 2

[11] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018. 1

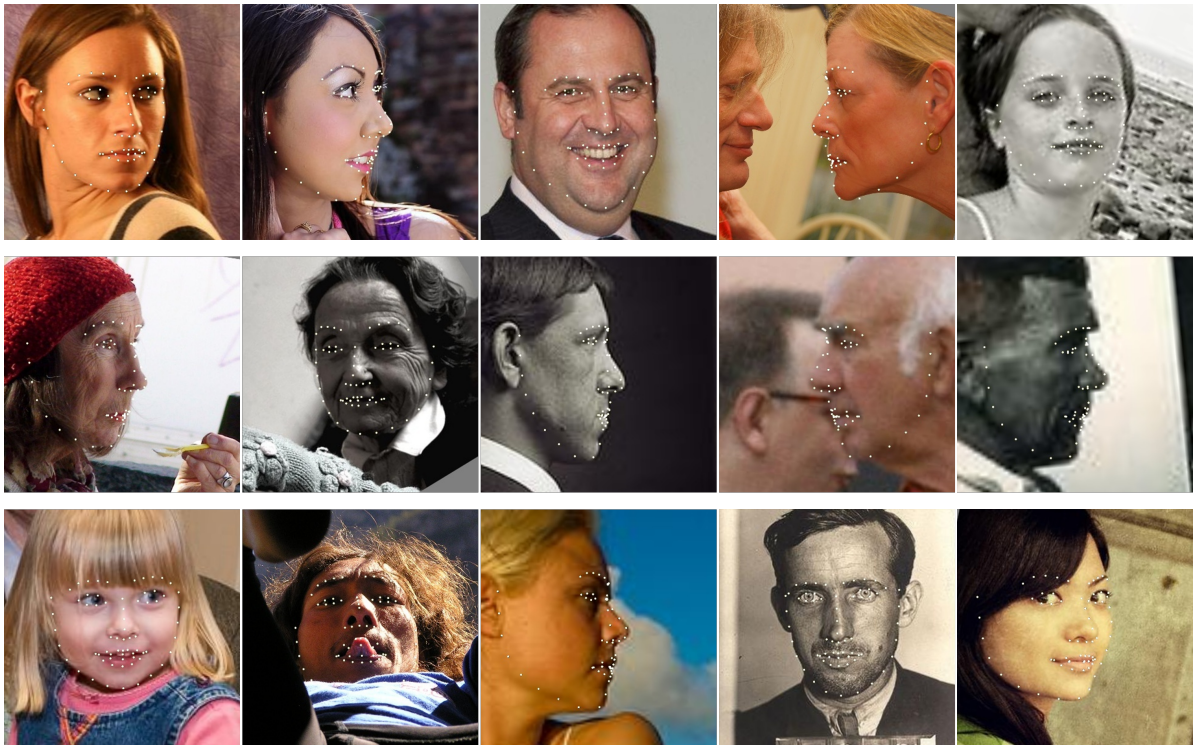

Figure 3: Sampled annotated images in the proposed AFLW 68-point dataset, including in-the-wild faces under large pose variations