Marginal Contrastive Correspondence for Guided Image Generation (Supplementary Material)

Fangneng Zhan
Nanyang Technological University

1. Appendix Outline

This supplementary document presents more details and experimental results which include: 2. Detailed Architecture, 3. Implementation Details, 4. More Ablation Study, 5. Limitations, 6. Ethical Considerations, and 7. More qualitative results, respectively.

2. Detailed Architecture

The architecture of the generation network in MCL-Net is consistent with CoCosNet [25]. The detailed architectures of the generator and discriminator in the generation network are shown in Fig. 1 and Fig. 2, respectively. The detailed architecture of the feature encoder in the alignment network is shown in Fig. 3.

3. Implementation Details

Due to the superior generation capability, there are numerous GAN-based image-to-image translation methods [9, 11–15, 19, 22, 26] that have been extensively investigated and achieved remarkable progress on translating different conditions such as semantic segmentation [1, 9, 10, 18, 21], key points [6, 7, 17, 20] and edge maps [2, 16, 27].

For the training setting and hyper-parameters, including learning rate, optimizer, etc., we follow the setting of CoCosNet for fair comparison. In detail, Adam solver with $\beta_1 = 0$ and $\beta_2 = 0.999$ is adopted for optimization. All experiments were conducted on 4 32GB Tesla V100 GPUs with synchronized BatchNorm. The default size for building correspondence is 64×64. The size of generated images is 256×256 in all generation tasks.

For the contrastive learning, we apply a two-layer MLP with 256 units at each layer to embed the encoder's features. We normalize the vector by its L2 norm. The temperature τ is 0.07 by default. Consistent with CUT [8], a small projection head (i.e., a two-layer MLP) is included to embed the encoded features.

Methods	FID	Style Relevance		Semantic
		Color	Texture	Consistency
w/o $\mathcal{L}_{cyc}$	28.73	**0.986**	0.962	0.863
w/o $\mathcal{L}_{fcst}$	29.69	0.988	**0.968**	0.867
w/o $\mathcal{L}_{per}$	46.32	0.972	0.876	0.817
w/o $\mathcal{L}_{cxt}$	38.04	0.963	0.945	0.864
w/o $\mathcal{L}_{pse}$	25.88	0.980	0.962	0.884
w/o $\mathcal{L}_{mcl}$	26.12	0.977	0.965	0.853
w/o $\mathcal{L}_{adv}$			5	0.853
Full Losses	**24.35**	0.984	0.967	**0.886**

Table 1. Ablation studies of different loss terms in MCL-Net over ADE20K [24] dataset.

4. More Ablation Study

We follow the setting of CoCosNet [23], except including the proposed marginal contrastive loss and self-correlation map. We performed several ablation studies to examine the contribution of each loss by removing it from the overall objective. Table 1 show experimental results over the dataset ADE20K. We can see that all involved losses contribute to the image translation performance in different manners and amounts.

5. Limitations and Future Work

The proposed method incorporates the self-correlation map for building correspondence. The learning of self-correlation map is driven by the proposed maginal contrastive learning. However, the learned self-correlation map is still not accurate enough, e.g., missing some structure information. We would explore employing separate contrastive learning to learn the self-correlation map implicitly, or using pre-trained model to extract self-correlation map directly in our future work.

6. Ethical Considerations:

This work aims to synthesize high-fidelity images with given conditional inputs and exemplar images. It could have negative impacts if it is used in illegal applications such as image forgery.

7. More Qualitative Results

We provide more conditional translation results with different exemplars on three tasks as shown in Figs. 4, 6, 5.

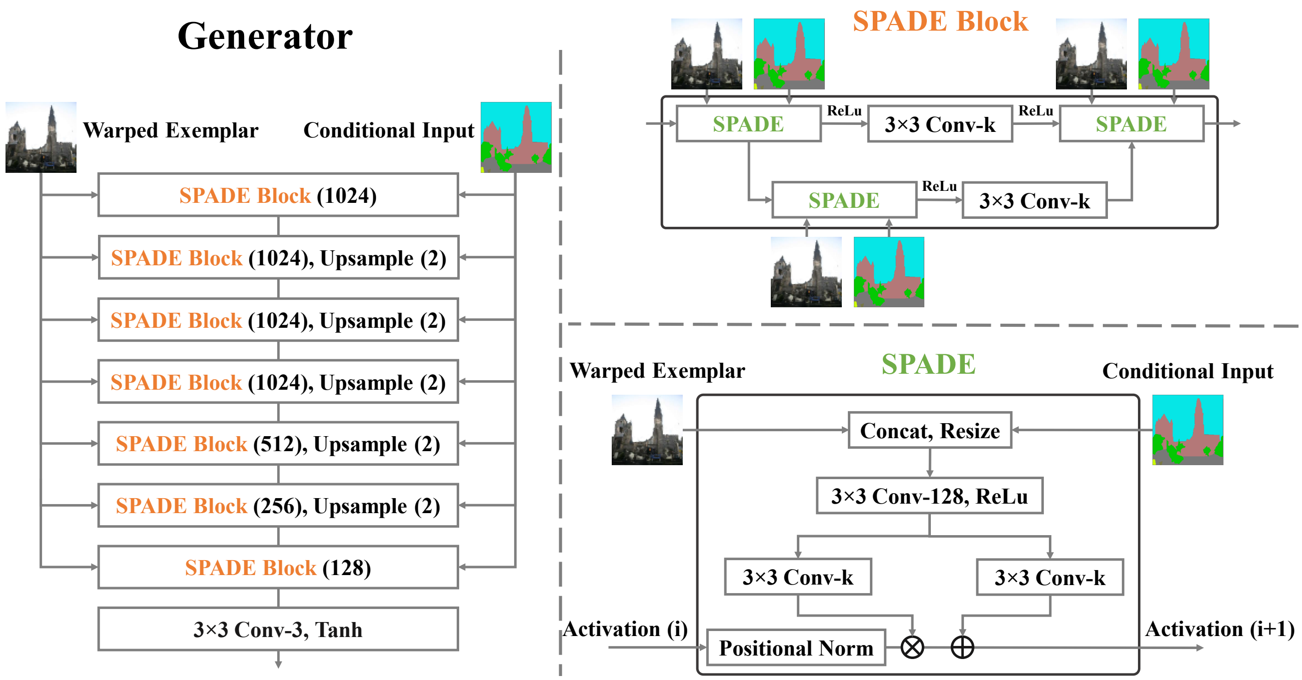

Figure 1. The structures of the *Generator* in our generation network: Positional Norm denotes positional normalization [3].

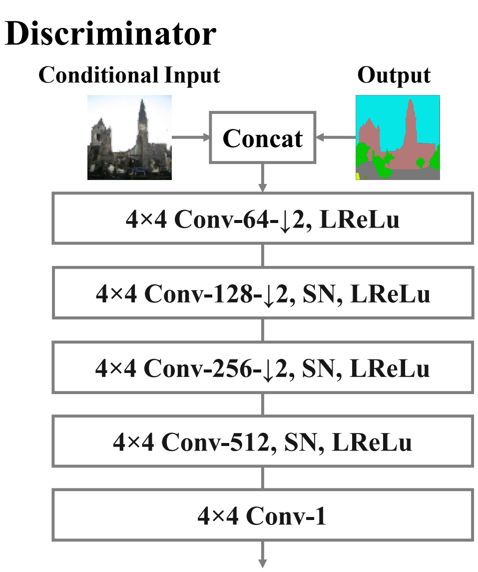

Figure 2. The structures of the *Discriminator* in our generation network: SN denotes spectrum normalization.

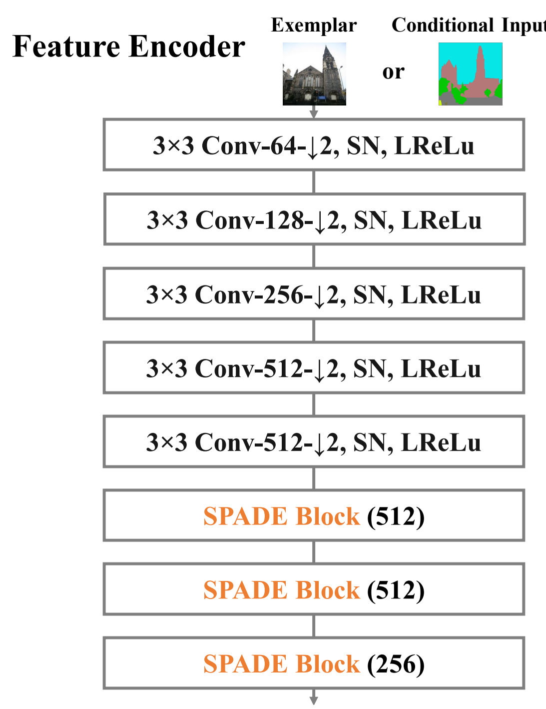

Figure 3. The structures of the *Feature Extractor* in our correspondence network.

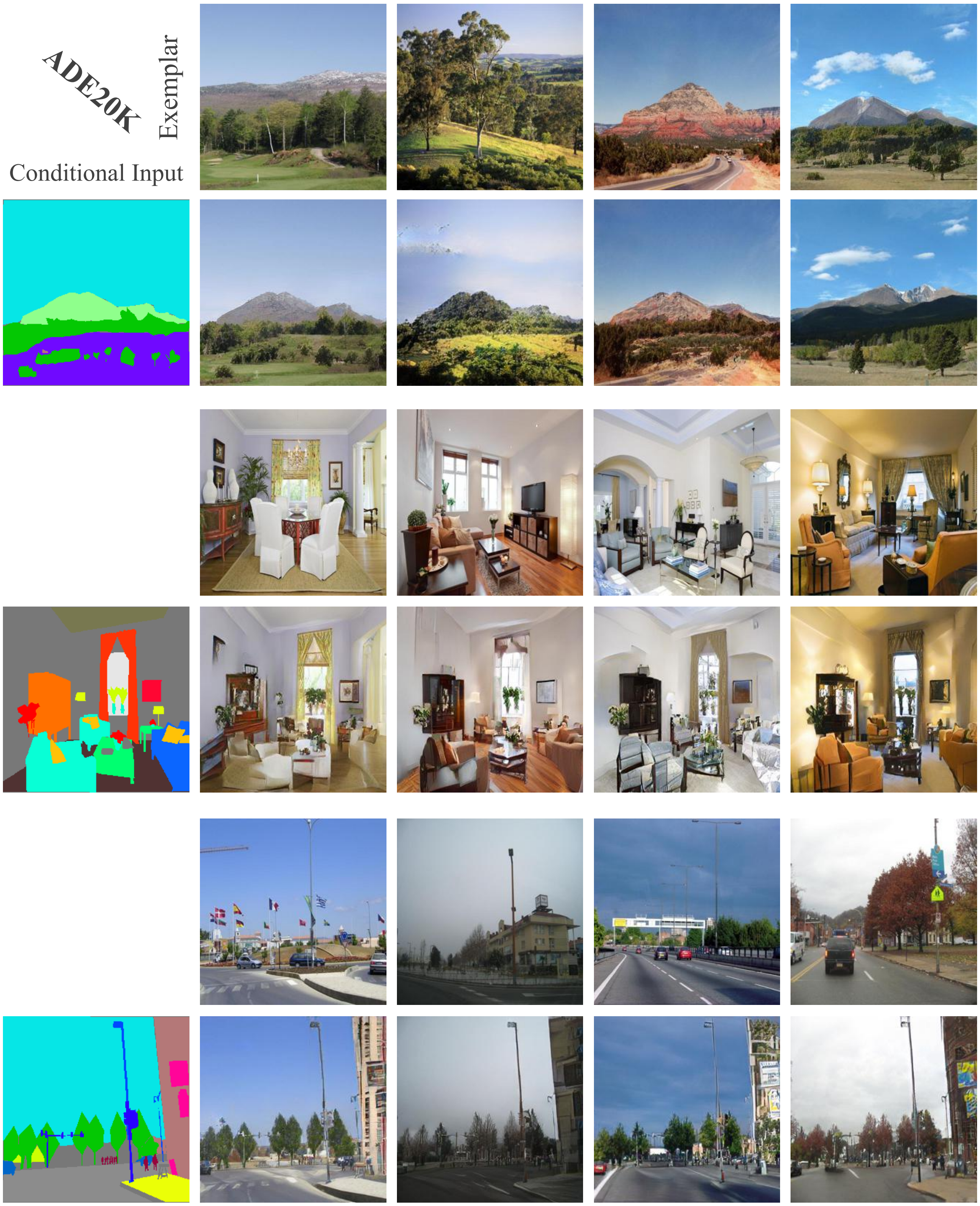

Figure 4. MCL-Net image generation from semantic maps over ADE20k [24].

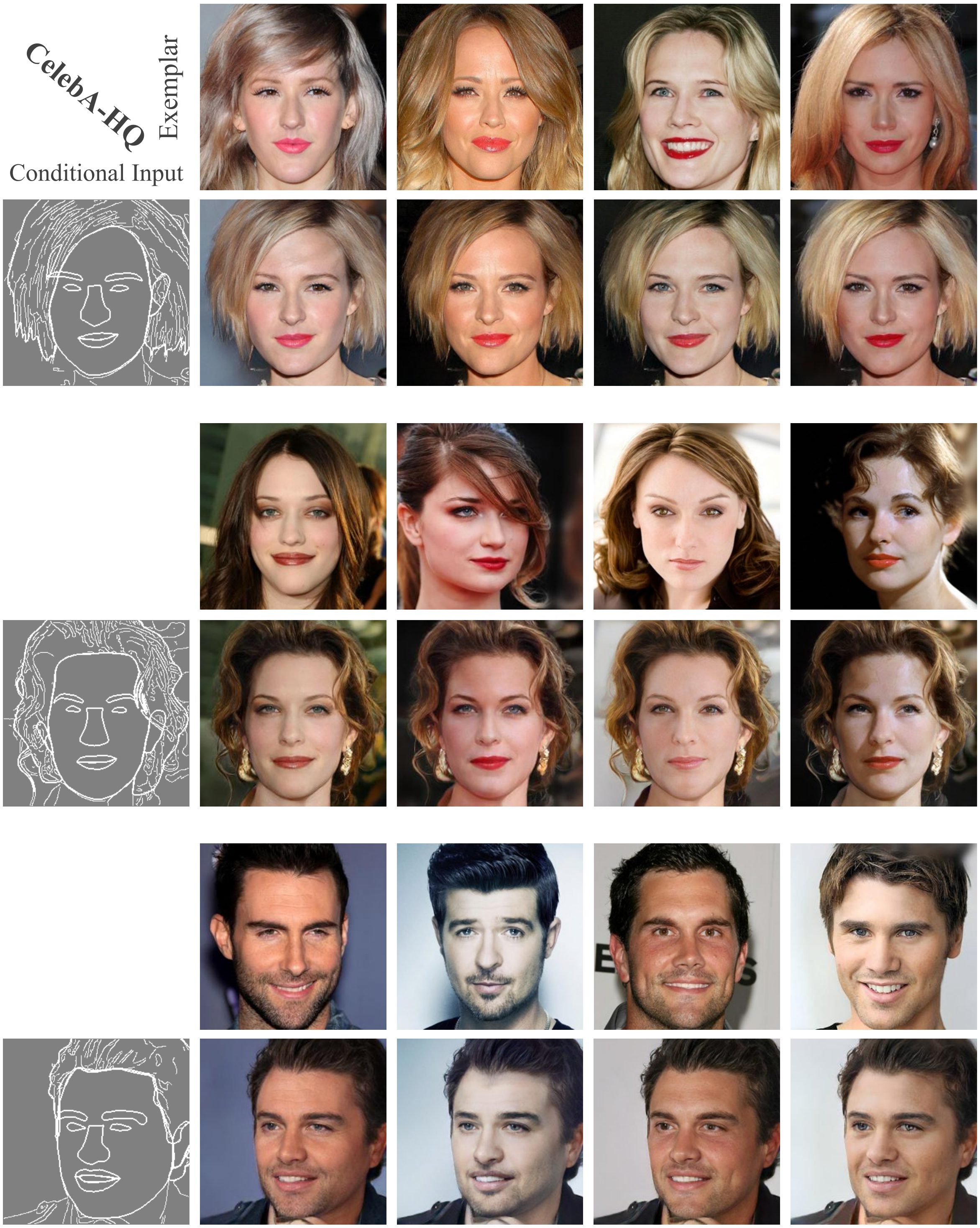

Figure 5. MCL-Net image generation from edge maps over CelebA-HQ [5].

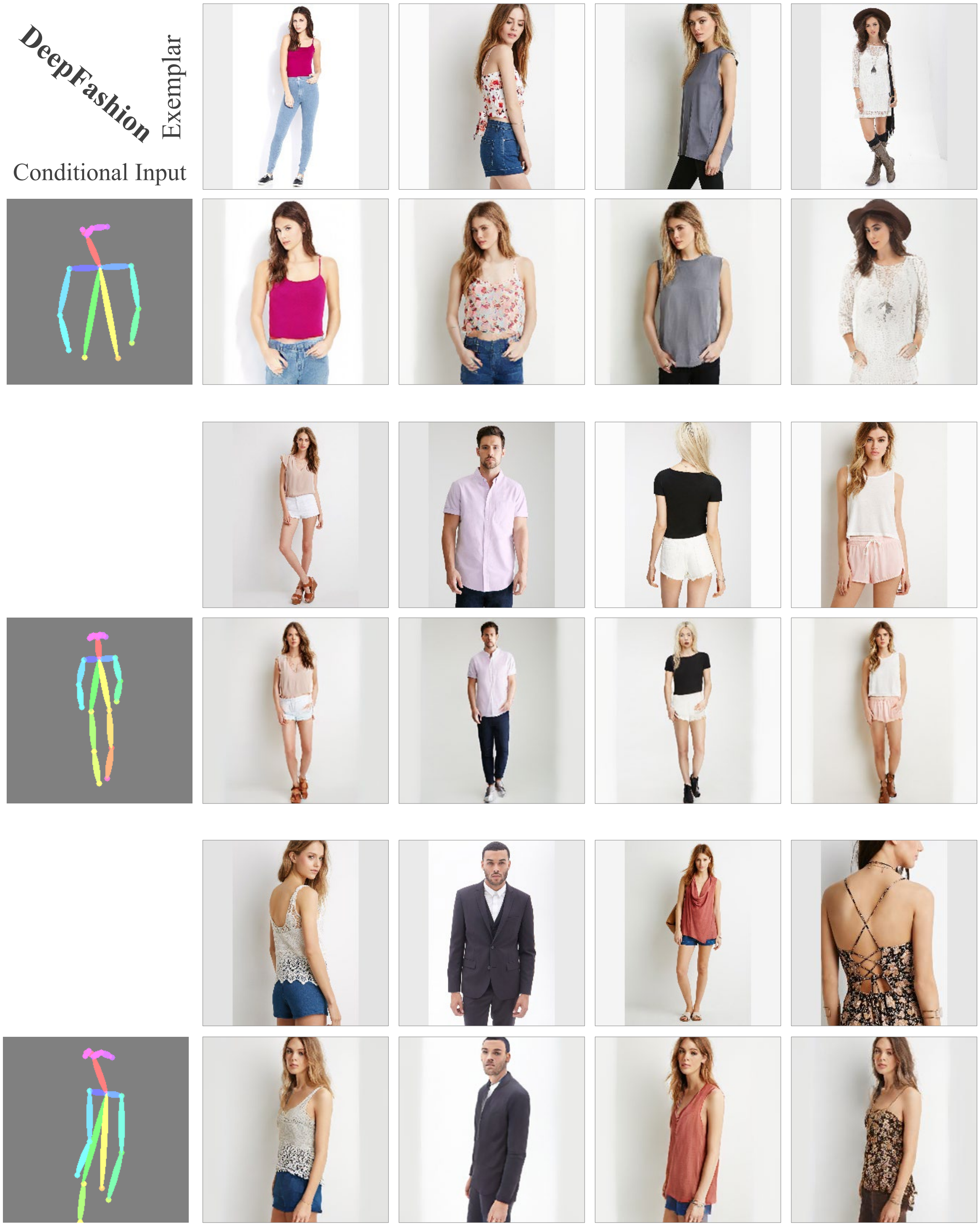

Figure 6. MCL-Net image generation from key points over dataset DeepFashion [4].

References

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[2] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 1

[3] Boyi Li, Felix Wu, Kilian Q Weinberger, and Serge Belongie. Positional normalization. In *Advances in Neural Information Processing Systems*, pages 1622–1634, 2019. 3

[4] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 6

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 5

[6] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in neural information processing systems*, pages 406–416, 2017. 1

[7] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 1

[8] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 1

[9] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 1

[10] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 1

[11] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill: A wavelet-based generation network for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14114–14123, 2021. 1

[12] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 69–78, 2021. 1

[13] Fangneng Zhan, Shijian Lu, Changgong Zhang, Feiying Ma, and Xuansong Xie. Adversarial image composition with auxiliary illumination. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1

[14] Fangneng Zhan, Chuhui Xue, and Shijian Lu. Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9105–9115, 2019. 1

[15] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[16] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Kaiwen Cui, Aoran Xiao, Shijian Lu, and Ling Shao. Bi-level feature alignment for versatile image translation and manipulation. *arXiv preprint arXiv:2107.03021*, 2021. 1

[17] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Changgong Zhang, Shijian Lu, Ling Shao, Feiying Ma, and Xuansong Xie. Gmlight: Lighting estimation via geometric distribution approximation. *arXiv preprint arXiv:2102.10244*, 2021. 1

[18] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, and Shijian Lu. Multimodal image synthesis and editing: A survey. *arXiv preprint arXiv:2112.13592*, 2021. 1

[19] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, and Changgong Zhang. Marginal contrastive correspondence for guided image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10663–10672, 2022. 1

[20] Fangneng Zhan, Changgong Zhang, Yingchen Yu, Yuan Chang, Shijian Lu, Feiying Ma, and Xuansong Xie. Emlight: Lighting estimation via spherical distribution approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3287–3295, 2021. 1

[21] Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Rongliang Wu, and Shijian Lu. Modulated contrast for versatile image synthesis. *arXiv preprint arXiv:2203.09333*, 2022. 1

[22] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3653–3662, 2019. 1

[23] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 1

[24] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1, 4

[25] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition, pages 11465–11475, 2021. 1

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1

[27] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017. 1