Unbalanced Feature Transport for Exemplar-based Image Translation: Supplementary Materials

Fangneng Zhan

1. Appendix Outline

This appendix presents more details and experimental results including literature comparison, user study, detailed network structure, limitation, and more qualitative illustration, respectively.

2. Literature Comparison

There are various methods for high-fidelity image synthesis [20, 20, 2, 16, 18, 28, 25, 38, 26, 30, 23, 24, 22, 29, 27, 21]. Existing works explored different conditional inputs such as semantic segmentation [6, 20, 16], scene layouts [17, 34, 10], key points [14, 15, 32], edge maps [6, 36, 8], etc. for photo-realistic image translation. On the other hand, optimal style control remains a critical yet challenging task that has attracted increasing attention in recent years. For example, [5] and [13] transfer style codes from exemplars to source images via adaptive instance normalization (AdaIN) [4]. [16] uses variational autoencoder (VAE) [7] to encode exemplars for image translation. [2] employs a style encoder for style consistency between exemplars and the translated images. Different from the aforementioned methods that adopt latent vectors for style control, [33] learns dense semantic correspondences between conditional inputs and exemplars for image translation. Similar ideas have been explored in other translation tasks such as image colorization [3, 31] that also employs exemplars to build up semantic correspondences. We compare UNITE with several state-of-the-art image-to-image translation methods including 1) Pix2pixHD [20]; 2) Pix2PixSC [19]; 3) StarGAN v2[2]; 4) SPADE [16]; 5) SelectionGAN [18]; 6) SMIS [38]; 7) SEAN [37]; and 8) CoCosNet [33].

The official implementations of Pixel2PixelHD, Pix2PixSC, StarGAN v2, SelectionGAN did not conduct all the four translation tasks (i.e. semantic map to image, layout to image, edge to image, key points to image). We therefore adapted the official implementation and re-trained models for the four translation tasks. SMIS and SEAN included experiments on ADE20K[35] for semantic map to image translation task, while the project page of SMIS does not provide the pre-trained encoder for style control. SEAN does not provide the pre-trained model on ADE20K. We thus adapted and re-trained SMIS and SEAN for the four translation tasks. CoCosNet provides the pre-trained models of three tasks (semantic map to image, edge to image, key points to image), we thus re-trained CoCosNet on COCO-Stuff for layout to image translation only.

3. User Study

We conducted a user study to evaluate the quality of images that are synthesized by different methods. Specially, 100 pairs of images generated by all compared methods are shown to 20 users who selected the image with the best visual quality. Fig. 1 shows experimental results over four image translation datasets. We can observe that the images generated by UNITE are much more realistic according to the user feedback.

4. Detailed Network Structure

The architecture of the neural projector is similar to SPADE [16]. The detailed architectures of the Generator, Discriminator and Feature Extractor in our UNITE are shown in Figs. 2, 3, and 4, respectively.

5. More Qualitative Results

We provide more conditional image translation illustrations that use different exemplars over four translation tasks as shown in Figs. 5, 6, respectively.

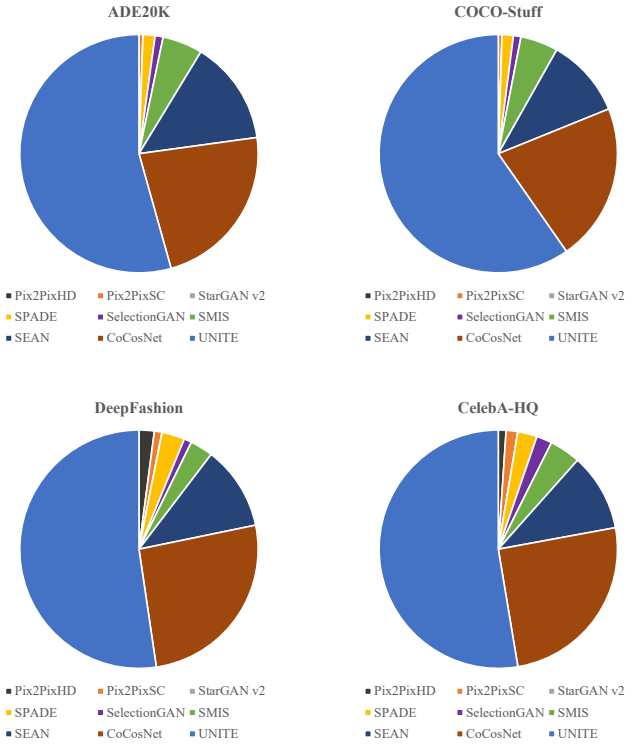

Figure 1. User study on four translation tasks on ADE20K [35], COCO-Stuff [1], DeepFashion [11], and CelebA-HQ [12].

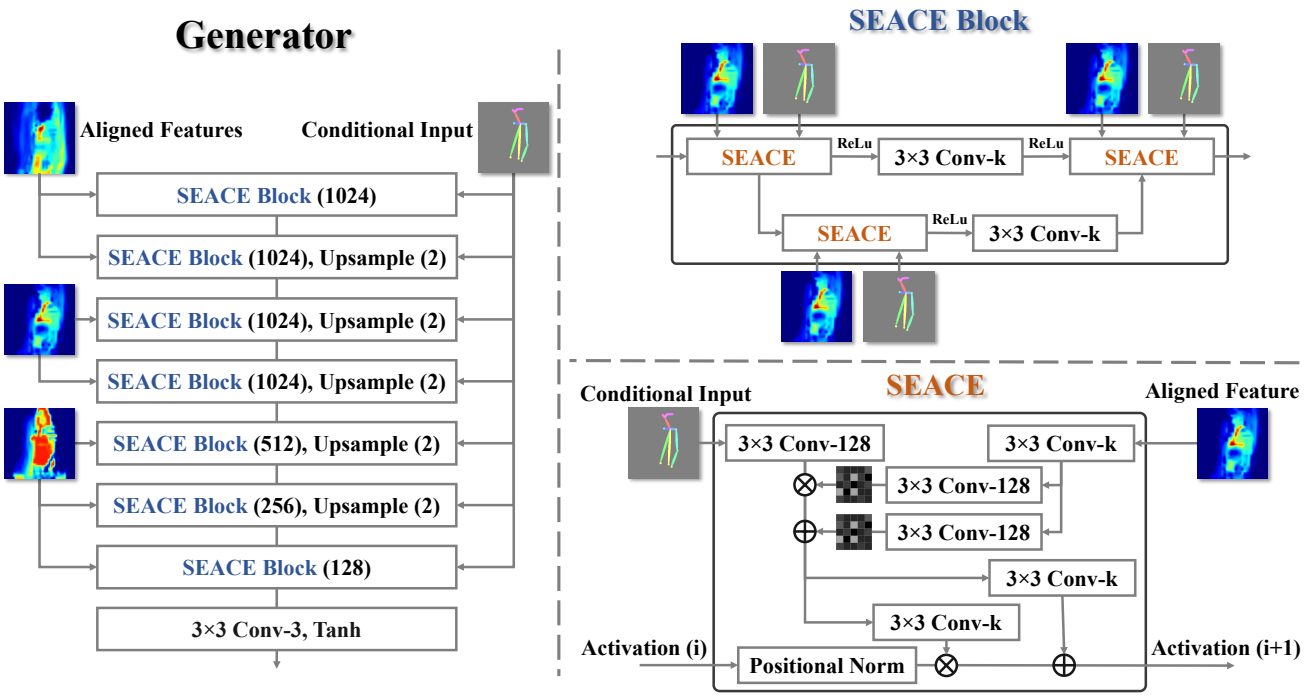

Figure 2. The structures of the *Generator* in our proposed image translation network: SEACE denotes the proposed semantic-activation normalization, and Positional Norm denotes positional normalization [9].

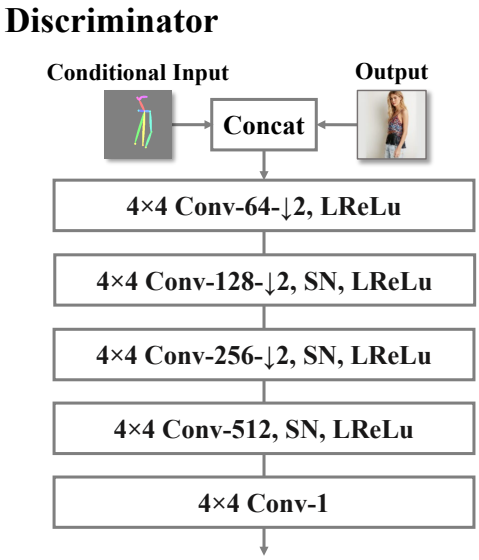

Figure 3. The structures of the *Discriminator* in our proposed image translation network: SN denotes spectrum normalization.

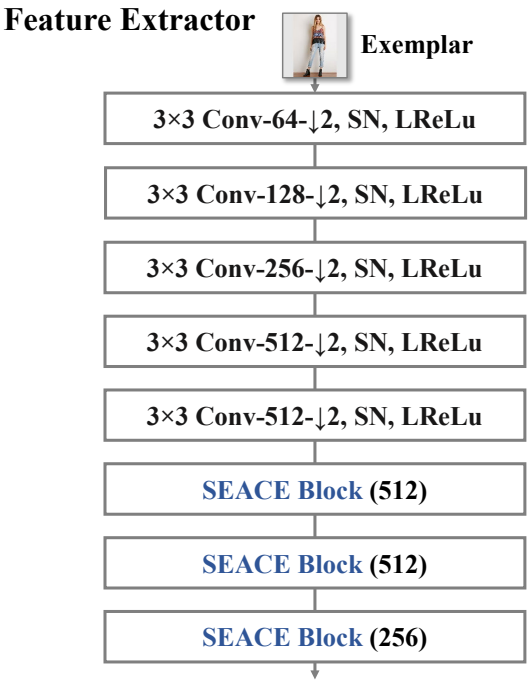

Figure 4. The structures of the *Feature Extractor* in our proposed feature transport network: The proposed SEACE is adopted in the last three layers to aggregate semantic feature.

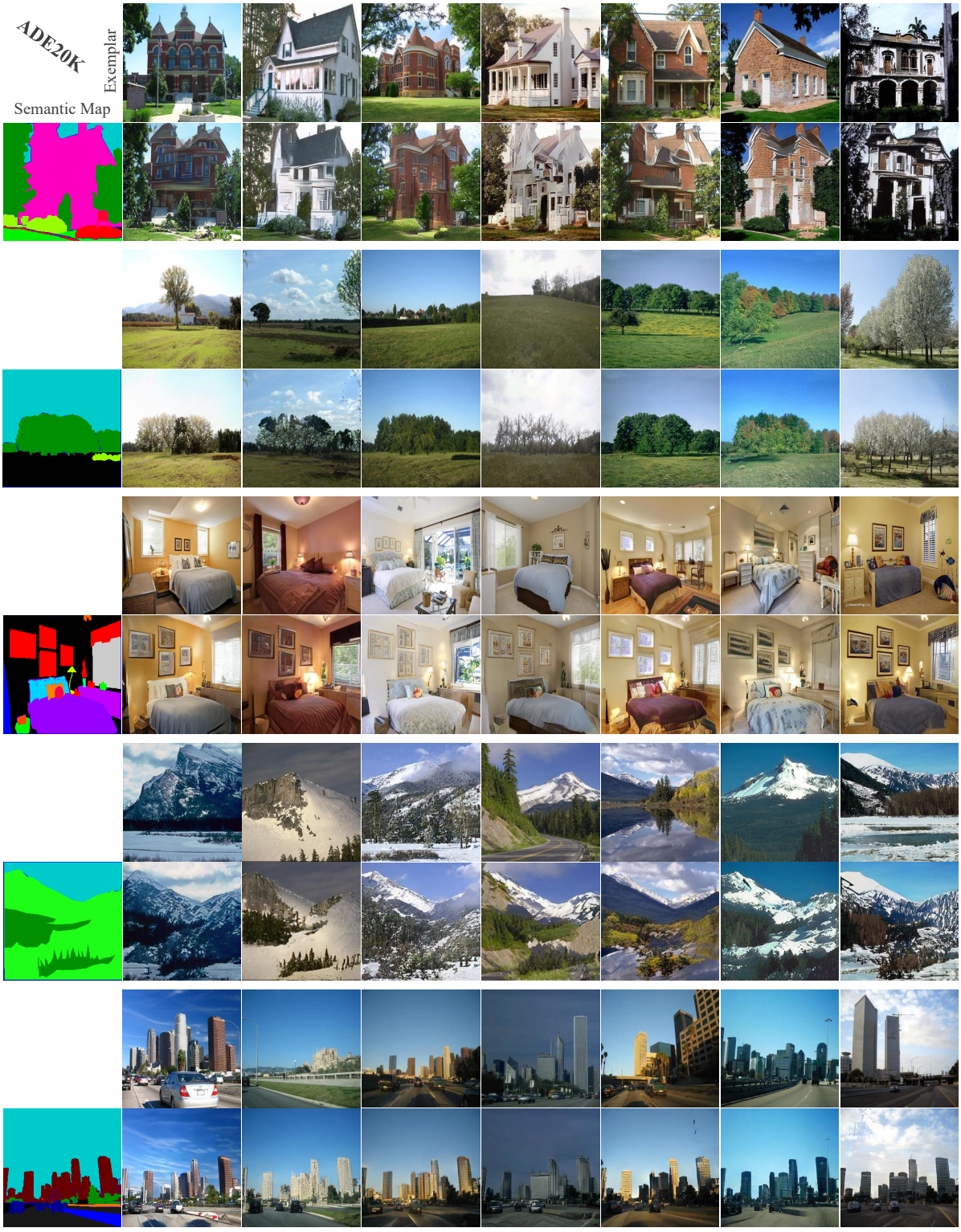

Figure 5. UNITE image generation from semantic maps over dataset ADE20K [35].

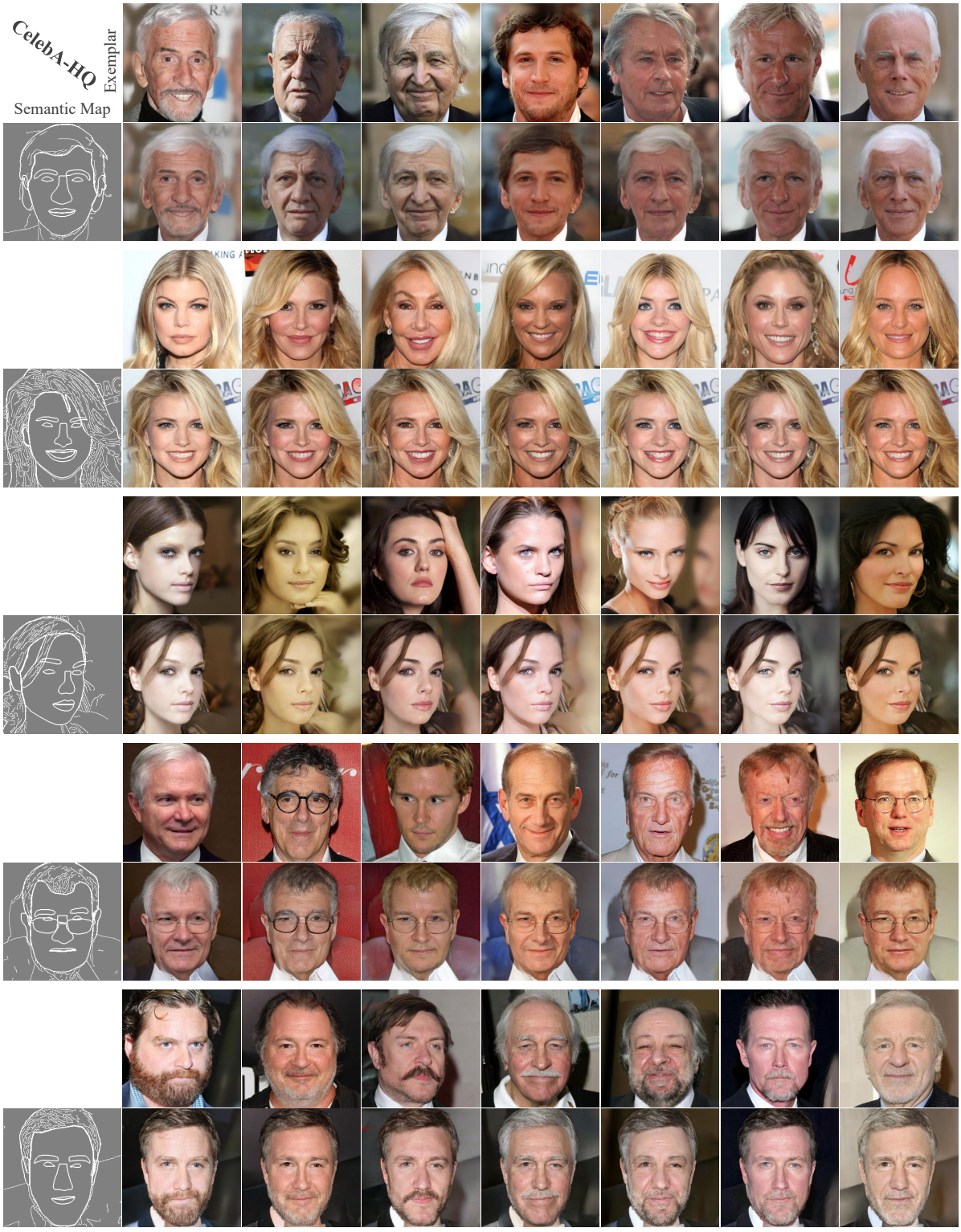

Figure 6. UNITE image generation from edge maps over CelebA-HQ [12].

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 2

[2] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 1

[3] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM Transactions on Graphics (TOG)*, 37(4):1–16, 2018. 1

[4] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 1

[5] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 1

[6] 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

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[8] 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

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

[10] Yandong Li, Yu Cheng, Zhe Gan, Licheng Yu, Liqiang Wang, and Jingjing Liu. Bachgan: High-resolution image synthesis from salient object layout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[11] 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. 2

[12] 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. 2, 5

[13] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. In *International Conference on Learning Representations*, 2018. 1

[14] 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

[15] 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

[16] 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

[17] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10531–10540, 2019. 1

[18] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2417–2426, 2019. 1

[19] Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter M Hall, and Shi-Min Hu. Example-guided style-consistent image synthesis from semantic labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1495–1504, 2019. 1

[20] 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

[21] Fangneng Zhan, Shijian Lu, and Chuhui Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–266, 2018. 1

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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

[27] 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

[28] 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

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

[30] 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

[31] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8061, 2019. 1

[32] Changgong Zhang, Fangneng Zhan, and Yuan Chang. Deep monocular 3d human pose estimation via cascaded dimension-lifting. *arXiv preprint arXiv:2104.03520*, 2021. 1

[33] 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

[34] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019. 1

[35] 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, 2, 4

[36] 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

[37] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 1

[38] Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. Semantically multi-modal image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5467–5476, 2020. 1