Modulated Contrast for Versatile Image Synthesis: Supplementary Material

Fangneng Zhan
Nanyang Technological University

This supplementary material presents more details and experimental results which include: 1. Pre-trained Segmentation for Evaluation, 2. Implementation Details, 3. More Analysis, 4. Limitations, 5. Ethical Considerations, and 6. More Qualitative Results, respectively.

1. Pre-trained Segmentation for Evaluation

We use pre-trained segmentation model to evaluate the quality of generated images conditioned on semantic maps. For Cityscapes dataset in unpaired image translation, DRN [15] with pre-trained DRN-D-105 model [1] is employed to evaluate the mean Average Precision (mAP), pixel Accuracy (pixAcc), and class Accuracy (classAcc).

For ADE20K dataset in paired image translation, the UPerNet [13] with pre-trained baseline-resnet101-upernet [2] [3] is adopted to evaluate the mean Intersection of Union (mIoU) and Accuracy (Acc).

2. Implementation Details

Multifarious image generation tasks [7–9, 18–20, 23, 25, 26, 28, 29] often entail multifaceted metrics to measure the inter-image similarity with regard to different properties such as image structures, image semantics and image perceptual realism, etc. There are various losses to achieve dedicated purposes in image synthesis [5, 6, 10–12, 16, 17, 21, 22, 24]. For instance, unpaired image translation is usually associated with certain losses to encourage correlation between the input and output images.

For the training setting of unpaired image translation, LSGAN loss [4], batch size of 12, Adam optimizer with learning rate of 0.002 are adopted for training. All models are trained up to 400 epochs for experiments on Cityscapes, Horse→Zebra, Winter→Summer. For the model architecture of unpaired image translation, we adopt the official implementation of CUT [4]. CycleLoss is implemented by adding a generator and discriminator. The selection of encoder layers and the corresponding weights of WeightNCE and MoNCE are consistent with PatchNCE [5], namely RGB pixels, the first and second down-sampling convolution, and the first and the fifth residual block. The receptive fields of the selected layers correspond to 1×1, 9×9, 15×15, 35×35, and 99×99. Experiments with F/LSeSim are based on the official implementation code [6].

For the training setting of paired image translation, we follow the hyper-parameter setting of SPADE [8], just replacing the perceptual loss with PatchNCE, and our WeightNCE, MoNCE. The model is trained up to 200, 60, and 100 epochs with a batchsize of 20 on ADE20K, CelebA-HQ, and DeepFashion datasets, respectively. For the model architecture of paired image translation, we adopt the official implementation of SPADE [7]. When applying PatchNCE, and our proposed WeightNCE and MoNCE on paired image translation, the selection of pre-trained VGG layers and the corresponding weights are consistent with the implementation of perceptual loss in SPADE, namely $relu1\_2, relu2\_2, relu3\_2, relu4\_2, relu5\_2$ layers with weights of 1/32, 1/16, 1/8, 1/4, 1.

In the contrastive learning, a two-layer MLP with 256 units at each layer is applied to embed the encoder's features which is further normalized through L2 norm. A temperature of 0.07 is adopted in contrastive learning which is consistent with CUT [7].

3. More Analysis

Our experiments show that a large negative term weight Q contributes to the FID score. However, the content preservation performance becomes worse with the increasing of Q as shown in Fig. 1. We conjecture that excessively large weight of negative term forces the contrastive learning to focus on the pushing of negative pairs and relatively ignore the pulling of positive pairs, thus degrading the contrastive learning performance.

4. Limitations and Future Work

The proposed method adjusts the weights of all negative samples. In fact, we aim to re-weight part of negative sam-

[1]https://github.com/fyu/drn
[2]https://github.com/CSAILVision/semantic-segmentation-pytorch
[3]http://sceneparsing.csail.mit.edu/model/pytorch/
[4]https://github.com/taesungp/contrastive-unpaired-translation
[5]https://github.com/taesungp/contrastive-unpaired-translation
[6]https://github.com/lyndonzheng/F-LSeSim
[7]https://github.com/NVlabs/SPADE

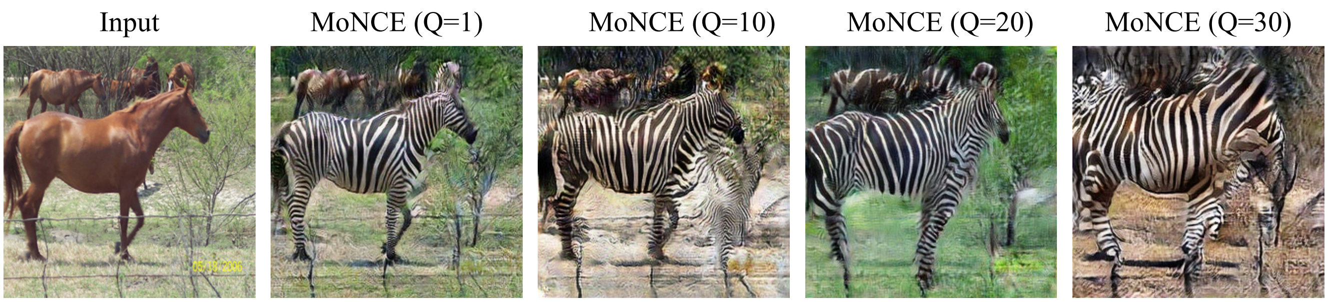

Figure 1. The unpaired image translation performance of MoNCE on Horse→Zebra with different negative term value Q. The default setting of MoNCE is $Q = 1$.

ples that may affect the contrastive learning significantly (negative or positive). Thus, certain threshold technique is expected to be employed to select part of the negative sample for fine re-weighting. On the other hand, differentiable top-k technique [14] enables to select elements in a differentiable way. In the future, we will explore differentiable top-k operation for the selection of negative sample for re-weighting.

5. Ethical Considerations:

The proposed method aims to boost the performance of image synthesis. It could have negative impacts if it is combined with other generation models for certain illegal purpose such as facilitating image forgery.

6. More Qualitative Results

We provide more image translation results including Figs. 2, 3, 4 for unpaired image translation, and Figs. 5, 6, 7 for paired image translation.

Figure 2. Qualitative comparison of different losses for unpaired image translation on Cityscapes (Semantic → Image) [1].

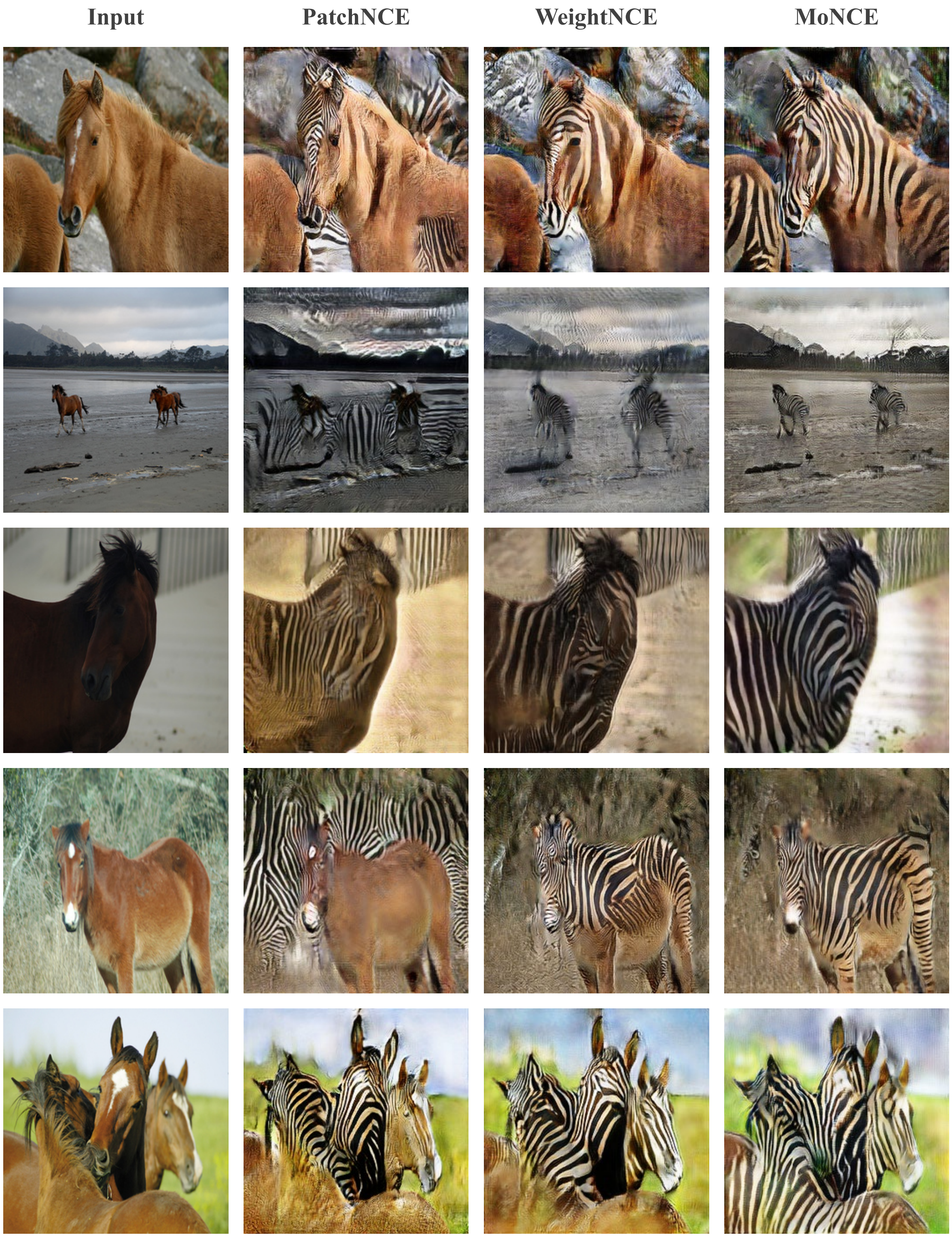

Figure 3. Qualitative comparison of different losses for unpaired image translation on Horse→Zebra [28].

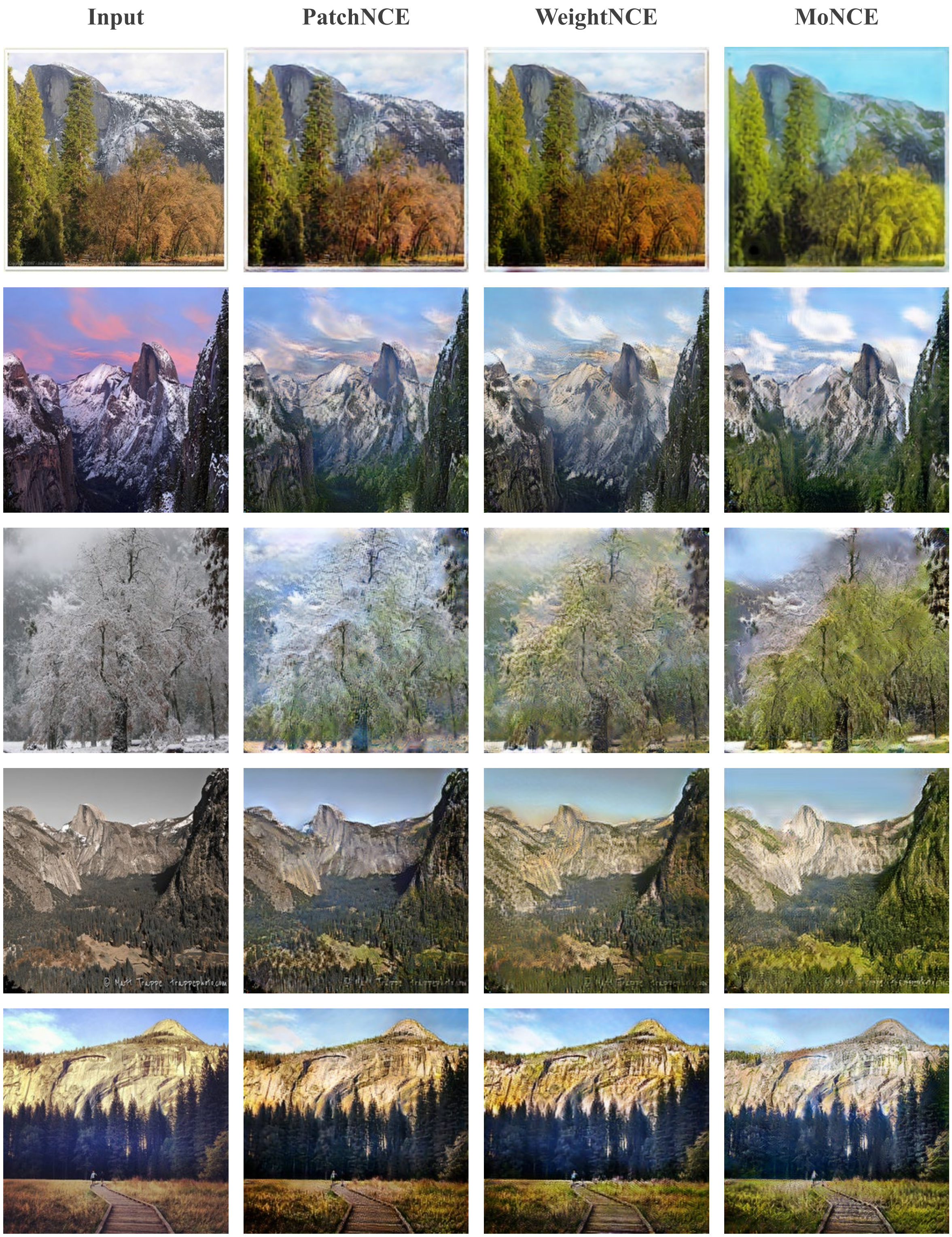

Figure 4. Qualitative comparison of different losses for unpaired image translation on Winter→Summer [28].

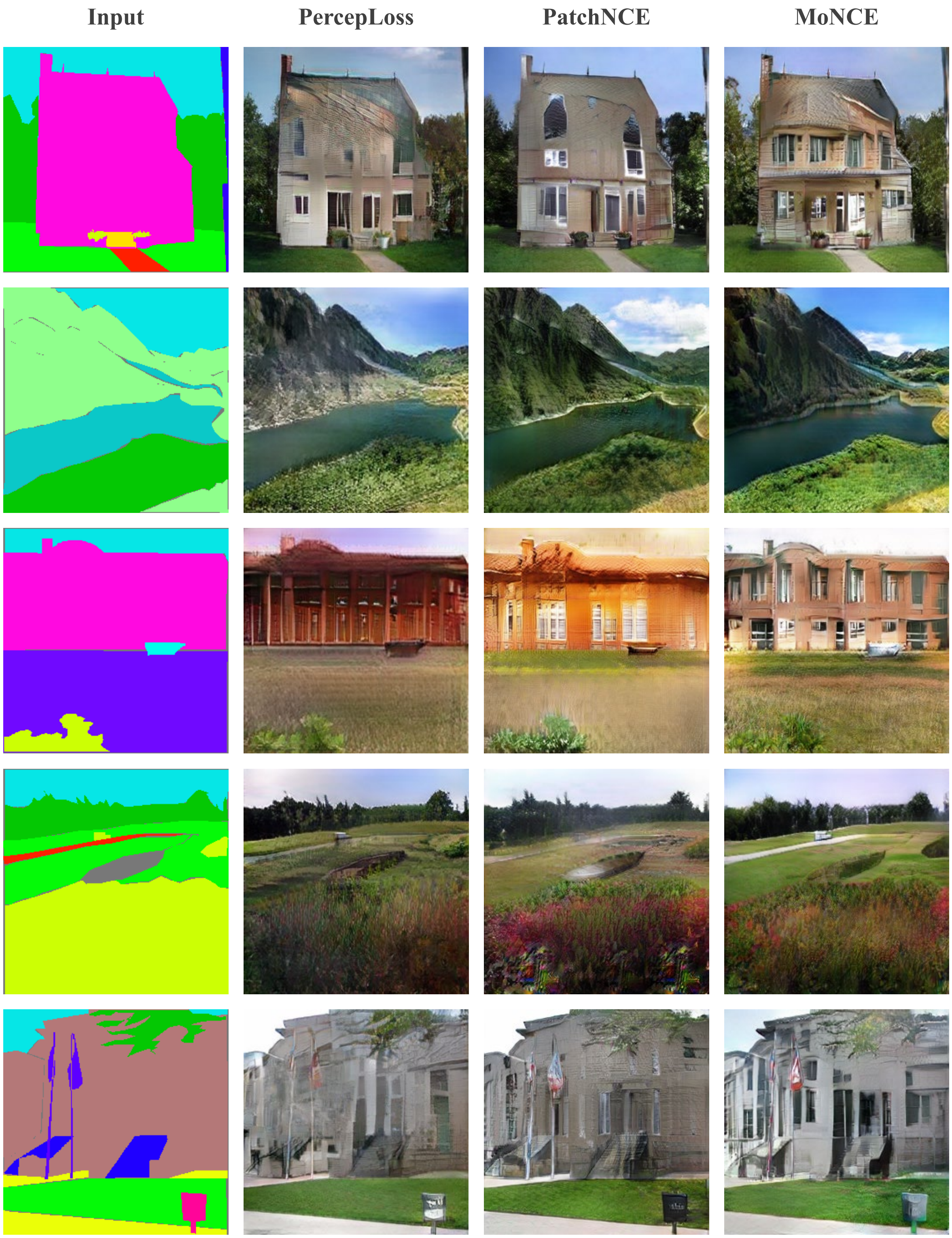

Figure 5. Qualitative comparison of different losses for paired image translation on ADE20K [27].

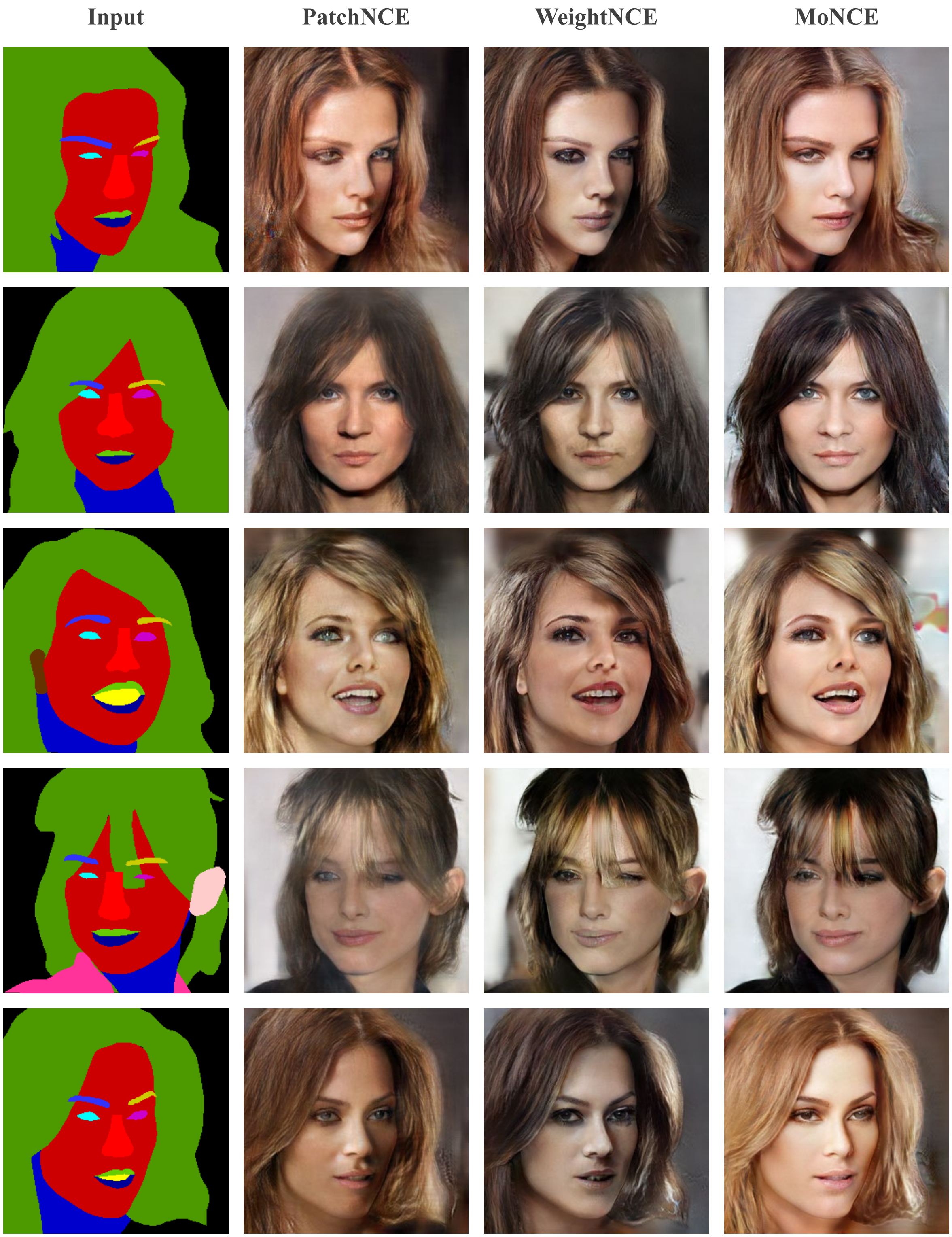

Figure 6. Qualitative comparison of different losses for paired image translation on CelebA-HQ (Semantic) [3].

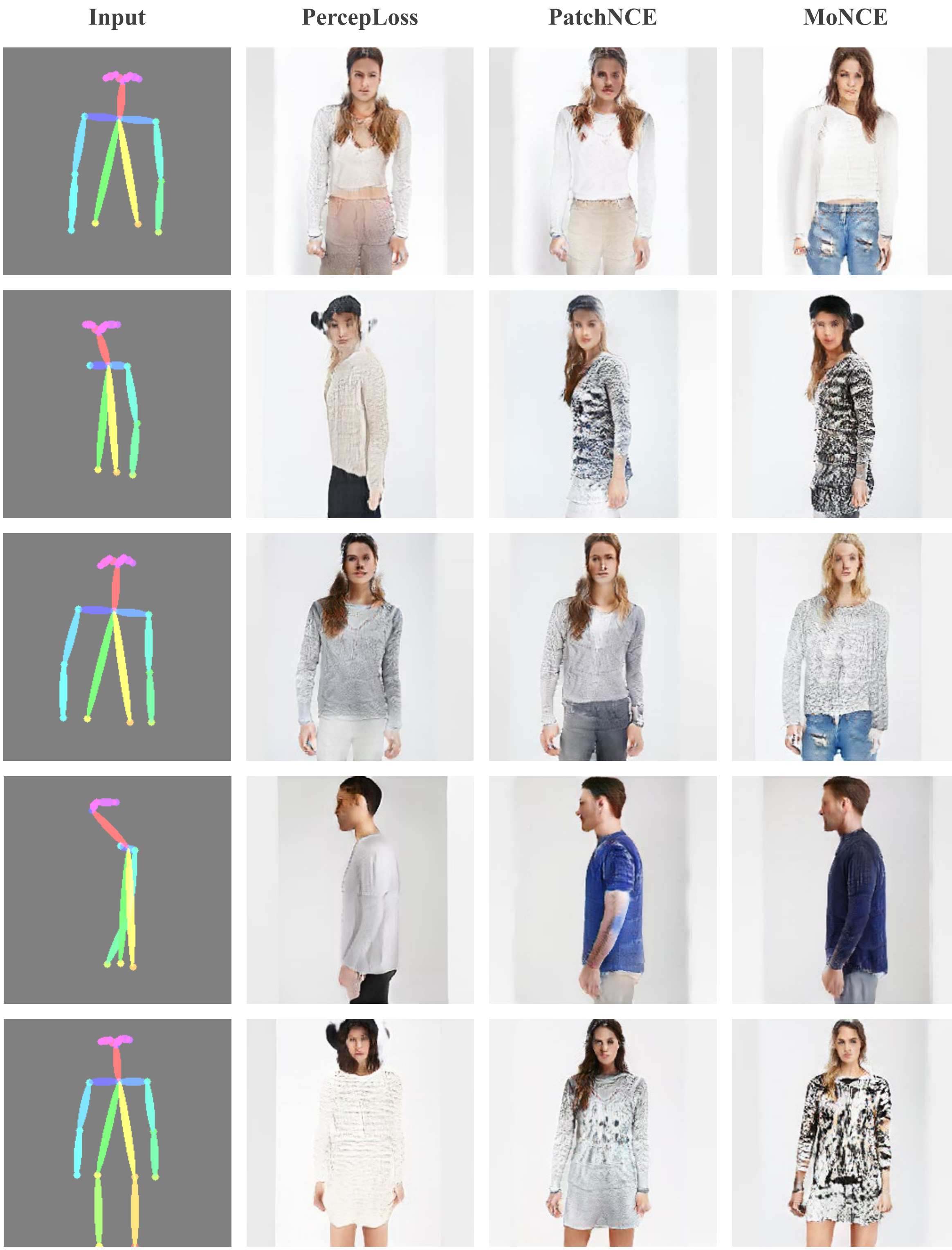

Figure 7. Qualitative comparison of different losses for paired image translation on DeepFashion (Keypoint) [2].

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[2] 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. 8

[3] 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. 7

[4] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 1

[5] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor. Maintaining natural image statistics with the contextual loss. In *Asian Conference on Computer Vision*, pages 427–443. Springer, 2018. 1

[6] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018. 1

[7] 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

[8] 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

[9] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017. 1

[10] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 1

[11] Rongliang Wu and Shijian Lu. Leed: Label-free expression editing via disentanglement. In *European Conference on Computer Vision*, pages 781–798. Springer, 2020. 1

[12] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade ef-gan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5021–5030, 2020. 1

[13] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 1

[14] Yujia Xie, Hanjun Dai, Minshuo Chen, Bo Dai, Tuo Zhao, Hongyuan Zha, Wei Wei, and Tomas Pfister. Differentiable top-k operator with optimal transport. *arXiv preprint arXiv:2002.06504*, 2020. 2

[15] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

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

[26] 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

[27] 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. 6

[28] 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, 4, 5

[29] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in neural information processing systems*, pages 465–476, 2017. 1