Supplemental Material for
CT-Net: Complementary Transfering Network for Garment Transfer with Arbitrary Geometric Changes

1. Network Structure

Let Ck denotes a Convolution layer with kernel size of 4, a stride 2, and k filters. Let Rk denotes the a Convolution layer with kernel size of 3, a stride 1, and k filters. Both of Ck and Rk are followed by InstanceNorm2d Normalization [6] and ReLu activation function. Let Lk denotes a Linear function output k dimension. Let ResBlock denotes the original Residual Block proposed in [3], in which the BatchNorm2d Normalization is replaced by InstanceNorm2d Normalization and the filters of the convolution layers are set as 256.

The two separate feature extractors $\mathcal{F}$ in Complementary Warping Module share the same structure: $\{$R64, C128, R256, C256, ResBlock $\times$ 6$\}$. We adopt the same regression net as [2] for the estimation of TPS warping, which consists of $\{$C512, C256, C128, C64, L32$\}$. Structure of generators in Layout Prediction Module and Dynamic Fusion Module are the same as U-Net [5], except the normalization is replaced by InstanceNorm2d Normalization. In Dynamic Fusion Module, we further employ one Convolution layer with kernel size of 3, a stride 1 at the end of the U-Net to estimate the attention mask. The discriminator is from pix2pixeHD [7].

2. Semantic Layout

We utilize the state-of-art LIP model [4] to estimate the human layouts, which contain 20 semantic labels to represent different human parts. LIP model makes a detailed prediction on the types of the clothes, including *upper clothes*, *dress*, *coat*, *pants*, *jumpsuits* and *skirt*, which makes the human parsing task much more challenge and involves extra noises (*e.g*. parts of the pants may be predicted as dress or skirt).

To simplify the task, we further merge the original parsing result to be a 7-channel human semantic layout, where each channel corresponds to *background*, *head*, *arms*, *legs*, *upper clothes*, *lower clothes* and *shoes*. In specific, we merge the *upper clothes*, *dress*, *coat* and *jumpsuits* in the original parsing results as the *upper clothes*; *pants* and *skirt* as the *lower clothes*.

We also apply the same merging process to the segmentation of densepose descriptor [1], leading to a 7-channel clothing-agnostic representations H^T.

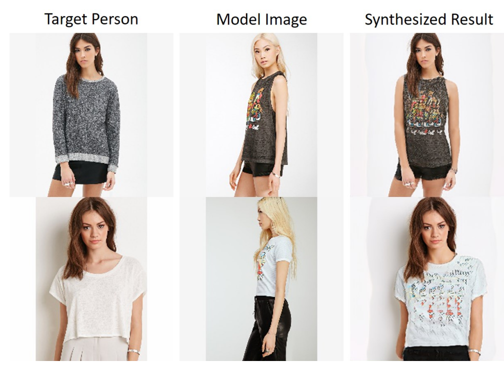

Figure 1. Failure cases when target view of the desired clothes has invisible regions in the clothes of the model image.

3. Limitations

The limitations of our model can be summarized as follows: (i) The performance of our network relys on the pretrained human parsing network. Our network fails to predict accurate target layouts and synthesize realistic garment transfer results when the parsing results are inaccurate, as shown in the last two rows of Figure 2. (ii) Our model is unaware of the 3D information of the clothes. When the target image has large clothing region not visible in the source image, the problem becomes ill-posed. Our model is only capable to utilize the warped visible regions to reconstruct the clothes in the target view, which may result in incorrect synthesized results as shown in Figure 1.

References

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 1

[2] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated

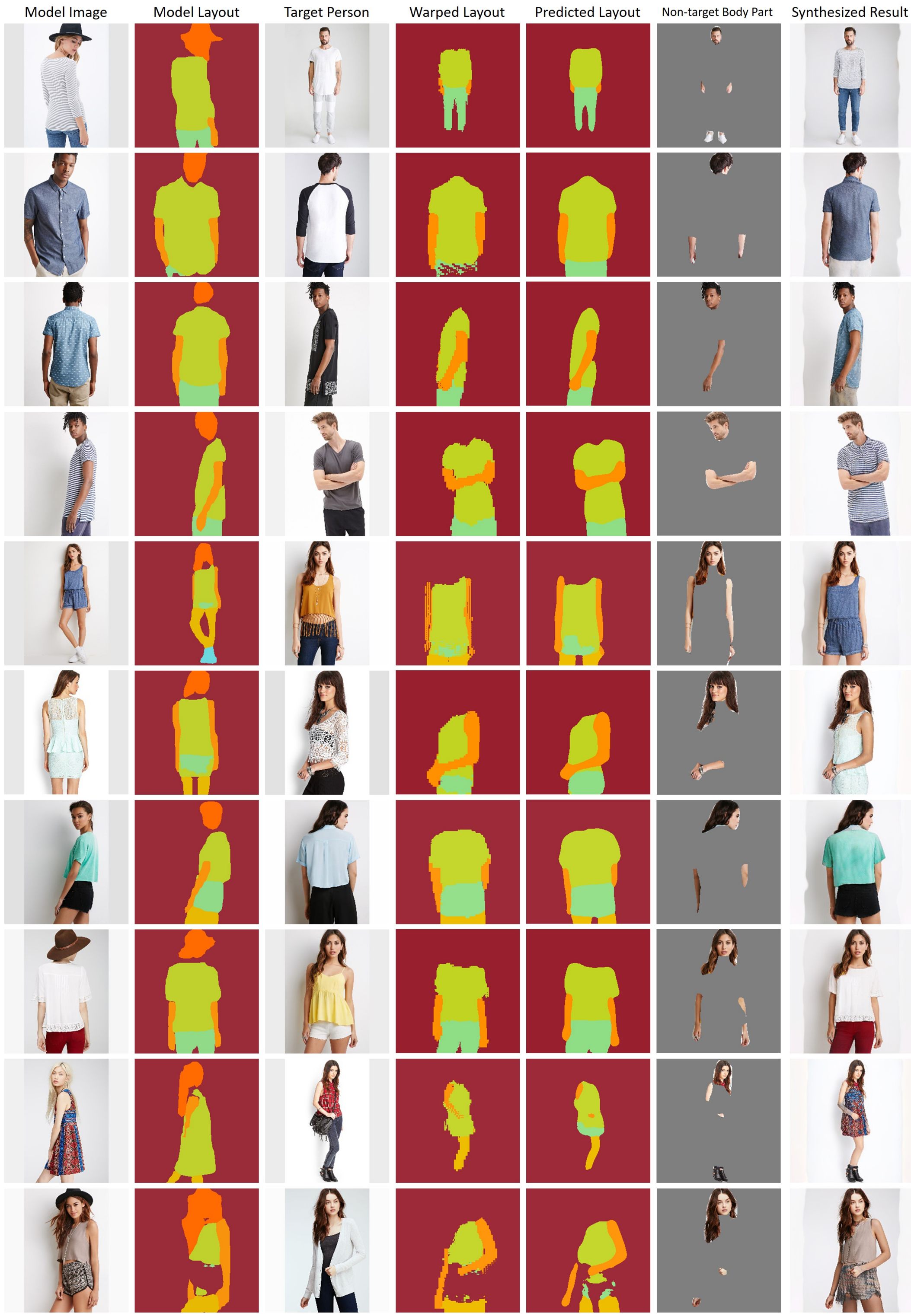

Figure 2. More visual results of the warped layout W^M, the predicted layout R^{lc} and the non-target body part I_u^T. The last two rows show two failure cases when the pretrained human parsing network fails to predict accurate parsing results.

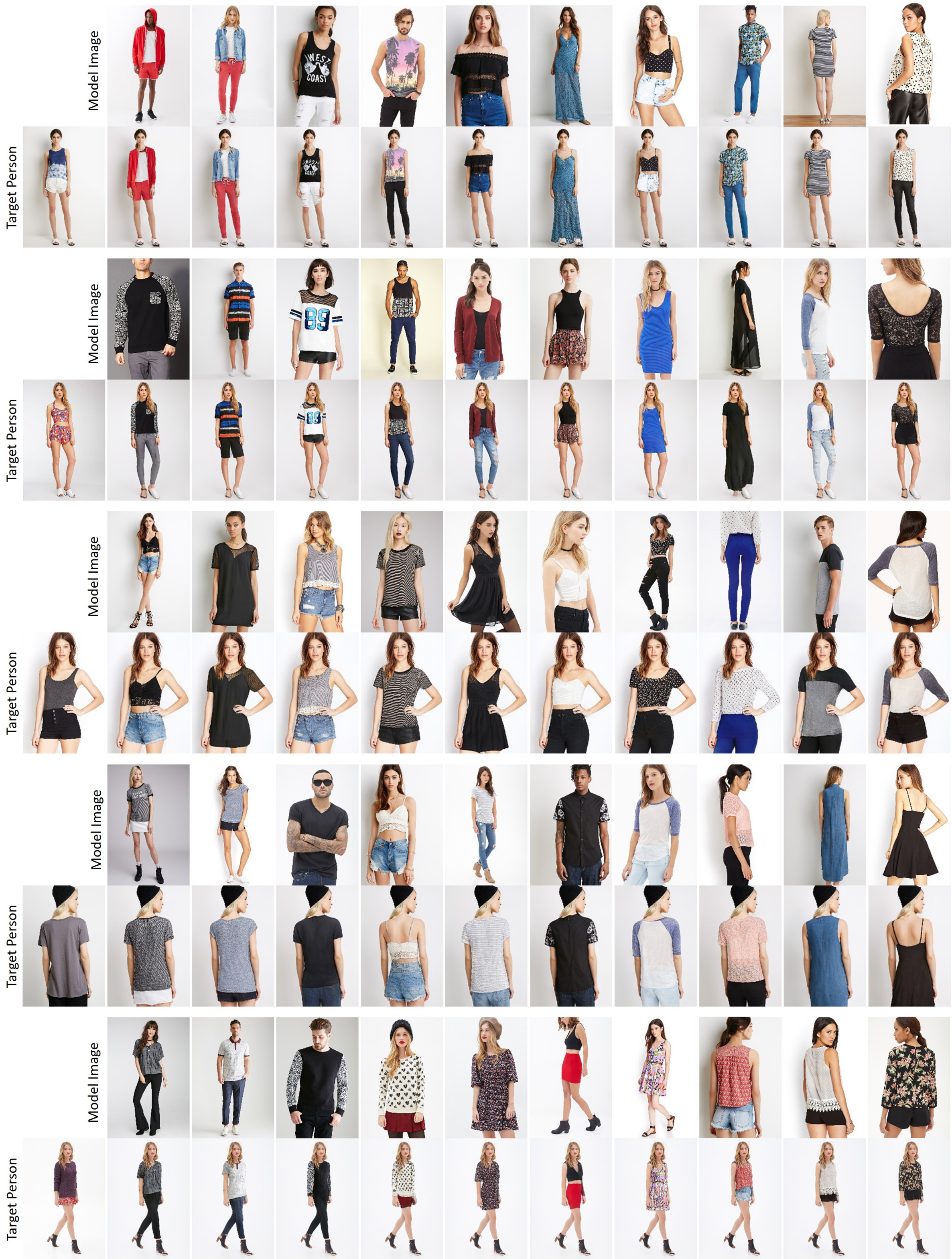

Figure 3. More garment transfer results generated by our CT-Net.

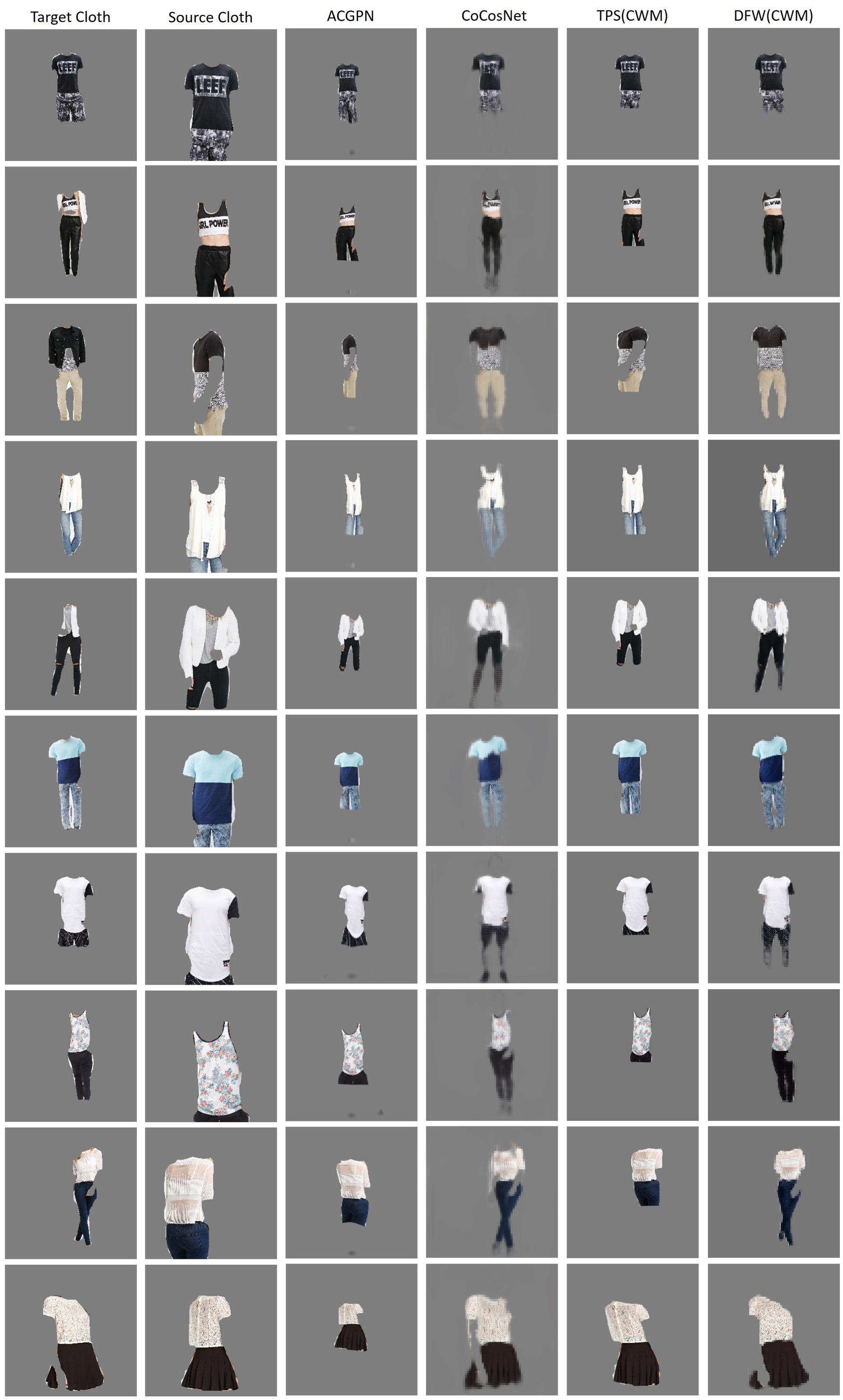

Figure 4. More visual comparisons of warping results. DFW(CWM) represents the warping results from DF-guided dense warping estimated in the Complementary Warping Module.

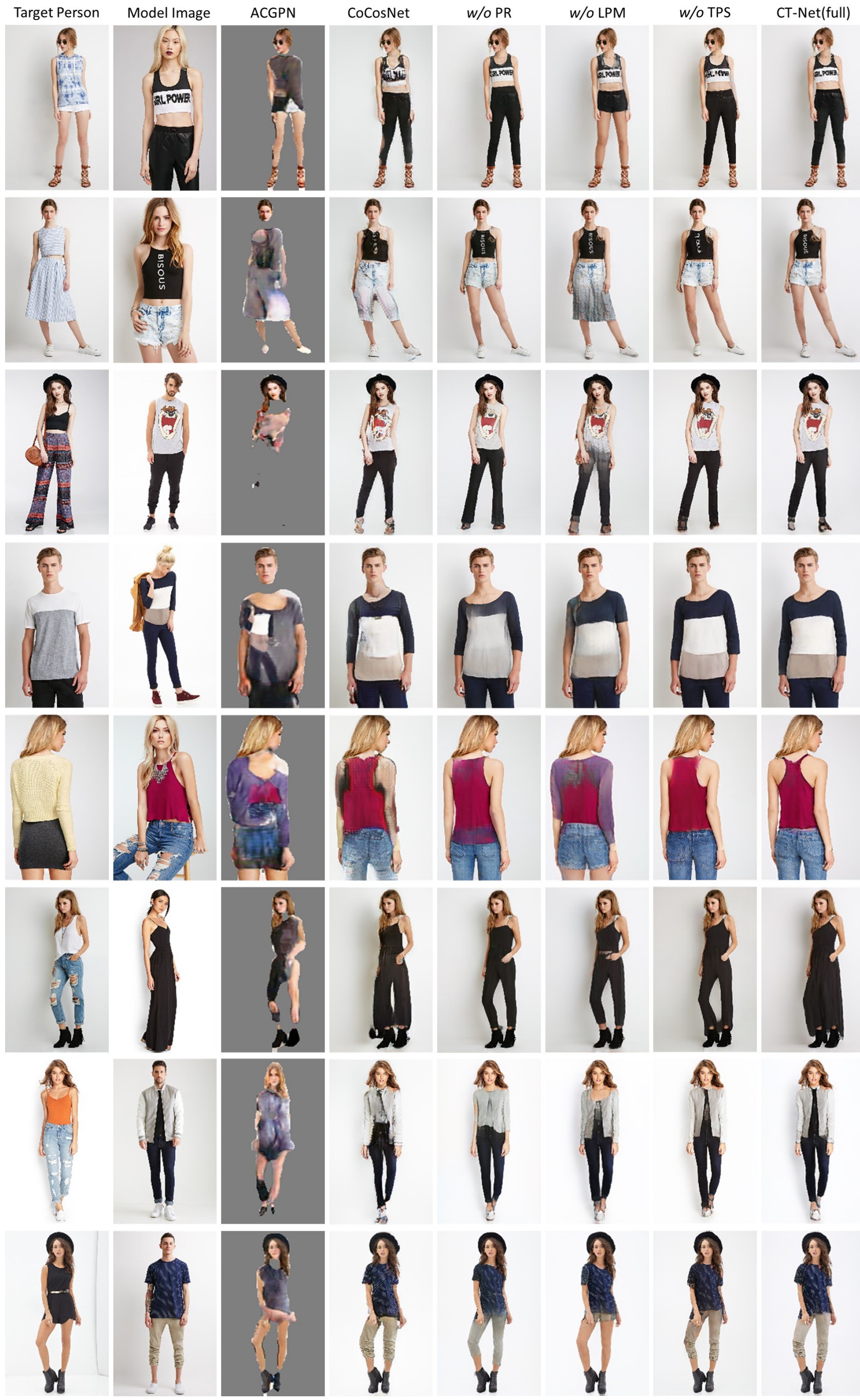

Figure 5. More qualitative comparisons with other methods.

warping gan for video virtual try-on. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1161–1170, 2019. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.

Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[4] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018. 1

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[6] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 1

[7] 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