REFERENCES

[1] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, ‘Viton: An image-based virtual try-on network’, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7543–7552.

[2] Z. Chong et al., ‘Catvton: Concatenation is all you need for virtual try-on with diffusion models’, arXiv preprint arXiv:2407. 15886, 2024.

[3] S. Choi, S. Park, M. Lee, and J. Choo, ‘Viton-hd: High-resolution virtual try-on via misalignment-aware normalization’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 14131–14140.

[4] E. J. Hu et al., ‘Lora: Low-rank adaptation of large language models’, ICLR, vol. 1, no. 2, p. 3, 2022.

[5] X. Li et al., ‘Warpdiffusion: Efficient diffusion model for high-fidelity virtual try-on’, arXiv preprint arXiv:2312. 03667, 2023.

[6] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, ‘Toward characteristic-preserving image-based virtual try-on network’, in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 589–604.

[7] M. R. Minar, T. T. Tuan, H. Ahn, P. Rosin, and Y.-K. Lai, ‘Cp-vton+: Clothing shape and texture preserving image-based virtual try-on’, in CVPR workshops, 2020, vol. 3, pp. 10–14.

[8] J. Kim, G. Gu, M. Park, S. Park, and J. Choo, ‘Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 8176–8185.

[9] Y. Xu, T. Gu, W. Chen, and A. Chen, ‘Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on’, in Proceedings of the AAAI Conference on Artificial Intelligence, 2025, vol. 39, pp. 8996–9004.

[10] D. Morelli, A. Baldrati, G. Cartella, M. Cornia, M. Bertini, and R.

Cucchiara, ‘Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on’, in Proceedings of the 31st ACM international conference on multimedia, 2023, pp. 8580–8589.

[11] P. Dhariwal and A. Nichol, ‘Diffusion models beat gans on image synthesis’, Advances in neural information processing systems, vol. 34, pp. 8780–8794, 2021.

[12] J. Ho, A. Jain, and P. Abbeel, ‘Denoising diffusion probabilistic models’, Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.

[13] J. Zhou et al., ‘Dream: Diffusion rectification and estimation-adaptive models’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8342–8351.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, ‘High-resolution image synthesis with latent diffusion models’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[15] Z. Xie, Z. Huang, F. Zhao, H. Dong, M. Kampffmeyer, and X. Liang, ‘Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan’, Advances in Neural Information Processing Systems, vol. 34, pp. 2598–2610, 2021.

[16] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, ‘Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 22500–22510.

[17] B. Yang et al., ‘Paint by example: Exemplar-based image editing with diffusion models’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 18381–18391.

[18] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, ‘Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models’, arXiv

preprint arXiv:2308. 06721, 2023.

[19] X. Chen, L. Huang, Y. Liu, Y. Shen, D. Zhao, and H. Zhao, ‘Anydoor: Zero-shot object-level image customization’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 6593–6602.

[20] L. Zhang, A. Rao, and M. Agrawala, ‘Adding conditional control to text-to-image diffusion models’, in Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 3836–3847.

[21] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, ‘Repaint: Inpainting using denoising diffusion probabilistic models’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11461–11471.

[22] I. Loshchilov and F. Hutter, ‘Decoupled weight decay regularization’, arXiv preprint arXiv:1711. 05101, 2017.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, ‘Image quality assessment: from error visibility to structural similarity’, IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.

[24] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, ‘The unreasonable effectiveness of deep features as a perceptual metric’, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.

[25] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0.

[26] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, ‘Demystifying mmd gans’, arXiv preprint arXiv:1801. 01401, 2018.

[27] A. Ramesh et al., ‘Zero-shot text-to-image generation’, in International conference on machine learning, 2021, pp. 8821–8831.

[28] C. Saharia et al., ‘Photorealistic text-to-image diffusion models with

deep language understanding', Advances in neural information processing systems, vol. 35, pp. 36479–36494, 2022.

[29] P. Esser, R. Rombach, and B. Ommer, 'Taming transformers for high-resolution image synthesis', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12873–12883.

[30] A. Q. Nichol and P. Dhariwal, 'Improved denoising diffusion probabilistic models', in International conference on machine learning, 2021, pp. 8162–8171.

[31] B. AlBahar, J. Lu, J. Yang, Z. Shu, E. Shechtman, and J.-B. Huang, 'Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan', ACM Transactions on Graphics (TOG), vol. 40, no. 6, pp. 1–11, 2021.

[32] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, 'Hierarchical text-conditional image generation with clip latents', arXiv preprint arXiv:2204. 06125, vol. 1, no. 2, p. 3, 2022.

[33] A. Radford et al., 'Learning transferable visual models from natural language supervision', in International conference on machine learning, 2021, pp. 8748–8763.

[34] O. Ronneberger, P. Fischer, and T. Brox, 'U-net: Convolutional networks for biomedical image segmentation', in Medical image computing and computer-assisted intervention--MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, 2015, pp. 234–241.

[35] X. Wang, S. Fu, Q. Huang, W. He, and H. Jiang, 'Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance', arXiv preprint arXiv:2406. 07209, 2024.

[36] T. Karras, S. Laine, and T. Aila, 'A style-based generator architecture for generative adversarial networks', in Proceedings of the IEEE/CVF

conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[37] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, ‘Image super-resolution via iterative refinement’, IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 4, pp. 4713–4726, 2022.

[38] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, ‘Image inpainting for irregular holes using partial convolutions’, in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 85–100.

[39] Z. Xue et al., ‘Raphael: Text-to-image generation via large mixture of diffusion paths’, Advances in Neural Information Processing Systems, vol. 36, pp. 41693–41706, 2023.

[40] X. Zhang et al., ‘Mmtryon: Multi-modal multi-reference control for high-quality fashion generation’, arXiv preprint arXiv:2405. 00448, 2024.

[41] S. Li et al., "Image Synthesis With Transformer-Based Diffusion Models," arXiv:2301.10972, 2023.

[42] X. Han, X. Hu, W. Huang, and M. R. Scott, ‘Clothflow: A flow-based model for clothed person generation’, in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 10471–10480.

[43] R. A. Güler, N. Neverova, and I. Kokkinos, ‘Densepose: Dense human pose estimation in the wild’, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7297–7306.

[44] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, ‘Context encoders: Feature learning by inpainting’, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536–2544.

[45] L. Zhu et al., ‘Tryondiffusion: A tale of two unets’, in Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4606–4615.

[46] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, ‘Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing’, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 932–940.

[47] R. Wang et al., ‘Stablegarment: Garment-centric generation via stable diffusion’, arXiv preprint arXiv:2403. 10783, 2024.

[48] X. Yang, C. Ding, Z. Hong, J. Huang, J. Tao, and X. Xu, ‘Texture-preserving diffusion models for high-fidelity virtual try-on’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 7017–7026.

[49] H. Wang, Z. Zhang, D. Di, S. Zhang, and W. Zuo, ‘Mv-vton: Multi-view virtual try-on with diffusion models’, in Proceedings of the AAAI Conference on Artificial Intelligence, 2025, vol. 39, pp. 7682–7690.

[50] M. Oquab et al., ‘Dinov2: Learning robust visual features without supervision’, arXiv preprint arXiv:2304. 07193, 2023.

[51] F. L. Bookstein, ‘Principal warps: thin-plate splines and the decomposition of deformations’, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 6, pp. 567–585, 1989.

[52] T. Brooks, A. Holynski, and A. A. Efros, ‘Instructpix2pix: Learning to follow image editing instructions’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 18392–18402.

[53] Y. Choi, S. Kwak, K. Lee, H. Choi, and J. Shin, ‘Improving diffusion models for authentic virtual try-on in the wild’, in European Conference on Computer Vision, 2024, pp. 206–235.

[54] A. Dosovitskiy et al., ‘An image is worth 16x16 words: Transformers

for image recognition at scale', arXiv preprint arXiv:2010. 11929, 2020.

[55] X. Liang et al., 'Deep human parsing with active template regression', IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 12, pp. 2402–2414, 2015.

[56] J. Gou, S. Sun, J. Zhang, J. Si, C. Qian, and L. Zhang, 'Taming the power of diffusion models for high-quality virtual try-on with appearance flow', in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 7599–7607.

[57] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep residual learning for image recognition', in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[58] J. Ho and T. Salimans, 'Classifier-free diffusion guidance', arXiv preprint arXiv:2207. 12598, 2022.

[59] D. P. Kingma, M. Welling, and Others, 'Auto-encoding variational bayes'. Banff, Canada, 2013.

[60] A. Vaswani et al., 'Attention is all you need', Advances in neural information processing systems, vol. 30, 2017.

[61] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo, 'Parser-free virtual try-on via distilling appearance flows', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8485–8493.

[62] Z. Xie et al., 'Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 23550–23559.

[63] Z. Xie et al., 'Was-vton: Warping architecture search for virtual try-on network', in Proceedings of the 29th ACM international conference on multimedia, 2021, pp. 3350–3359.

[64] Z. Xie et al., ‘Pasta-gan++: A versatile framework for high-resolution unpaired virtual try-on’, arXiv preprint arXiv:2207. 13475, 2022.

[65] Z. Xie, Z. Huang, F. Zhao, H. Dong, M. Kampffmeyer, and X. Liang, ‘Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan’, Advances in Neural Information Processing Systems, vol. 34, pp. 2598–2610, 2021.

[66] F. Zhao et al., ‘M3d-vton: A monocular-to-3d virtual try-on network’, in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 13239–13249.

[67] X. Ju, X. Liu, X. Wang, Y. Bian, Y. Shan, and Q. Xu, ‘Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion’, in European Conference on Computer Vision, 2024, pp. 150–168.

[68] P. Li, Y. Xu, Y. Wei, and Y. Yang, ‘Self-correction for human parsing’, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 3260–3271, 2020.

[69] N. Zhang and H. Tang, ‘Text-to-image synthesis: A decade survey’, arXiv preprint arXiv:2411. 16164, 2024.

[70] J. Qin, ‘Scalable Motion In-betweening via Diffusion and Physics-Based Character Adaptation’, arXiv preprint arXiv:2504. 09413, 2025.