Enhanced Realism in Virtual Try-On Tasks Using Diffusion Methods

Saris Kiattithapanayong
Department of Statistics
Faculty of Commerce and Accountancy, Chulalongkorn University
Bangkok, Thailand
saris.kiat@gmail.com

Suronapee Phoomvuthisarn
Department of Statistics
Faculty of Commerce and Accountancy, Chulalongkorn University
Bangkok, Thailand
suronapee@cbs.chula.ac.th

Abstract—**Virtual try-on technology is revolutionizing online retail by enabling customers to visualize garments on their bodies before purchasing. Traditional methods, often based on Generative Adversarial Networks (GANs), face challenges such as misalignment and visual artifacts, especially in complex poses. We present a virtual try-on framework leveraging diffusion models to enhance realism, accuracy, and garment detail preservation. Our approach integrates Vector Quantized Variational Autoencoders (VQ-VAEs) for precise feature matching within a diffusion U-Net architecture. By adopting image-based conditioning with the CLIP image encoder, our system utilizes visual features directly from clothing images for more faithful garment representations. Additionally, an Additional Feature Preserving Block (ControlNet) maintains intricate details like textures and logos, addressing fine-grained garment fidelity challenges. Quantitative evaluation demonstrates our system's superior performance, achieving the best LPIPS of 0.082. We also achieve a Fréchet Inception Distance (FID) of 7.782 and Kernel Inception Distance (KID) of 1.53, indicating enhanced image quality and feature alignment. Although the Structural Similarity Index Measure (SSIM) of 0.825 is slightly lower, it underscores the trade-off for improved realism and garment detail preservation. Our contributions set a new benchmark for accurate and realistic clothing visualization in virtual try-on systems.**

***Keywords*—Virtual Tryon, Diffusions, ControlNet**

I. Introduction

Virtual try-on technology has transformed the online fashion industry, allowing users to visualize how garments might look on their bodies, thereby enhancing customer experiences and reducing return rates. While substantial progress has been made with the development of image- based virtual try-on systems powered by Generative Adversarial Networks (GANs) [1, 2, 22], existing methods face critical challenges. These include difficulties in preserving fine-grained garment details such as logos and textures, generating natural and realistic body alignments, and handling variations in body posture.

Despite the promising capabilities of GAN-based methods, their limitations in producing high-resolution and artifact-free images have driven the exploration of alternative approaches. Diffusion-based models [3,4] have emerged as a compelling solution, leveraging iterative refinement to model complex data distributions and achieve superior image synthesis. Studies like LADI- VTON [10] demonstrate the potential of diffusion models for virtual try-on tasks. However, these methods still fall short in preserving intricate garment features and maintaining consistency in facial and body details, leaving significant gaps in generating lifelike, high-fidelity try-on images.

To address these gaps, this study introduces a virtual try-on framework that integrates latent diffusion models [3,4], Vector Quantized Variational Autoencoders (VQ-VAEs) [28], and an innovative image-based conditioning mechanism powered by the CLIP Vision Encoder [18]. Each of these components addresses specific challenges:

- VQ-VAE improves image synthesis quality by encoding images into a discrete latent space, capturing long-term dependencies for more accurate feature matching.
- CLIP Vision Encoder replaces traditional text- based conditioning with image-based conditioning, directly utilizing visual garment features for more precise and realistic try-on outputs.
- Feature Preserving Blocks are introduced within the diffusion model to retain intricate garment details, such as logos and textures, which are often lost in previous methods.

The framework is evaluated using a comprehensive set of metrics, including LPIPS. FID, KID, SSIM for perceptual similarity and qualitative assessments, against state-of-the-art models like VITON-HD [6], HR-VTON [22], TRYONDIFFUSION [19], and LADI-VTON [10]. The results demonstrate superior performance, with our method achieving the best LPIPS score (0.0820) FID score (7.782) and KID score (1.53), highlighting its ability to generate high-fidelity, realistic virtual try-on results.

By addressing critical challenges in garment detail preservation, alignment, and pose realism, this work not only advances the state of virtual try-on systems but also establishes a strong foundation for future innovations. In the following sections, we will discuss the proposed framework in detail, outlining the architecture, training methodology, and experimental evaluations.

II. Literature Review

A. Virtual Tryon

1) Virtual Try-On Using GANs

GAN-based approaches have tackled the virtual try-on problem through two stages: deforming clothing to fit the target body and blending it with the person's image [1, 2, 7, 12, 14, 29]. Early methods like VITON [1] and CP-VITON [2] used Thin Plate Spline (TPS) transformations in Geometric Matching Modules (GMM), but these struggled with large deformations like reshaping long sleeves. Later approaches, such as Appearance Flow [29], employed local pixel sampling for greater flexibility in handling complex deformations.

However, GAN-based models face persistent misalignment issues, leading to artifacts in reshaped clothing, especially with complex poses involving occlusions and deformations. Methods like VTNFP [7] and ACGPN [2] improved alignment by incorporating segmentation maps, but challenges remain in generating high-resolution images. Misalignments and reliance on simple U-Net architectures [12] limit image quality, particularly in synthesizing occluded body parts, while increasing computational costs.

2) Diffusion-Based Virtual Try-On

Denoising Diffusion Probabilistic Models (DDPM) [10, 19] generate realistic images by reversing a gradual noising process and have emerged as promising tools for virtual try-on. Tryon Diffusion [19] employs two U-Nets for normal diffusion and feature preservation but requires large, paired datasets, making it computationally intensive. Latent diffusion models (LDM) [17] address this by using VAEs to perform diffusion in latent space, reducing resource requirements. For example, LADI-VTON [10] employs a VAE for encoding, treating virtual try-on as a diffusion inpainting task [21]. It uses a CLIP-based pseudo-word encoder [18] with cross-attention to encode target clothing. However, this approach suffers from spatial information loss, limiting its ability to preserve high-frequency details like text and patterns, which are essential for real-world scenarios. Diffusion Model

3) Diffusion Model

Denoising Diffusion Probabilistic Models (DDPM) [4] have been introduced as a method for generating realistic images by reversing a progressive noise-adding process starting from a normal distribution. While capable of producing diverse and high-quality images, DDPM suffers from slow sampling speeds, limiting its practical use. Further optimizing efficiency, Latent Diffusion Models (LDM) [17] shifted the denoising process to a latent space using pre-trained encoder-decoder networks, reducing computational demands. With these advancements, diffusion models have become a strong alternative to GANs for generative tasks.

Researchers have also focused on enhancing control over diffusion-based generation. Text-to-image methods [17] incorporate textual input during the denoising phase, guiding models to produce text-aligned visuals. However, applying diffusion models to virtual try-on remains challenging. Text-to-image methods fall short in capturing the intricate visual details of garments, and inpainting approaches [21] struggle with precise control over modified areas. To overcome this, inspired by [26] we propose Additional Semantic Preserving Block into a diffusion model for refinement, leveraging fine-grained local conditions alongside global constraints during denoising. This approach ensures greater control and accuracy in rendering garment details.

III. Prepare Your Paper Before Styling

A. Model Overview

As illustrated in Fig. 1, the architecture of the proposed model processes an input image of a person and a clothing image through an autoencoder, a diffusion U- Net, and feature preservation blocks. The primary objective is to synthesize an image that seamlessly integrates the person's features with the clothing details.

This is achieved by employing advanced techniques, such as cross-attention mechanisms and latent diffusion, which enhance the quality of image generation. By preserving intricate garment details and accurately aligning clothing with the person's pose, the model delivers realistic and high-fidelity outputs.

Specifically tailored for virtual try-on applications, this framework allows users to visualize how they would look in various clothing items, offering an intuitive and immersive alternative to physical trials. This capability has significant implications for e-commerce, enabling improved customer experiences and reducing product return rates.

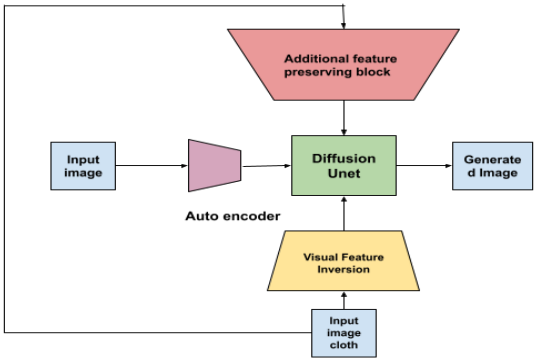

Fig. 1. Model Components Overview

B. Autoencoder (Latent Space Encoder)

The first component of the architecture, the autoencoder, transforms RGB input image $x \in R^{H \times W \times 3}$, the encoder E compresses x into a latent representation $z = E(x)$, and the decoder D reconstructs the image $\tilde{x} = D(z) = D + E(x)$ where $z \in R^{C \times H \times W}$. The encoder reduces the image dimensions by a factor $c \times \left(\frac{H}{f}\right) \times \left(\frac{W}{f}\right)$with experiments conducted using $f = 8$ To regularize the process, vector quantization regularization (VQ-reg.) is implemented by embedding a vector quantization layer [28] directly into the decoder, seamlessly incorporating quantization into the decoding process.

C. Visual Feature Inversion (CLIP Image Projection)

The second component of the architecture integrates detailed information about the target garment Ic into the diffusion model's conditioning mechanism via visual feature projection. CLIP aligns image and text inputs in a shared embedding space,

with the visual encoder extracting Ic 's features and mapping them into the CLIP token embedding space W. A visual inversion adapter $F\theta$ refines these embeddings to accurately represent Ic's characteristics, serving as key conditioning for the diffusion process. The inpainting region, defined by a mask M, fully covers the target garment for seamless replacement. By leveraging the visual semantics from CLIP and conditioning the Stable Diffusion network $\epsilon\theta$, the method achieves realistic and detailed garment rendering for virtual try-ons.

D. Diffusion UNET (Tryon)

An illustration of the Diffusion Tryon process, referred to as Training Loop 1, is shown in Fig. 2. Starting with a person image $I_p \in R^{H\times W\times D}$, a clothing-agnostic representation $I_{p\mathrm{m}}$, called the masked cloth-out person image, is generated by removing garment information from I_p. Virtual try-on is framed as an exemplar-based image inpainting task, where $I_{p\mathrm{m}}$ is completed using the clothing image I_{c}. The U-Net input combines: (1) a noisy latent image $Z(I_p)$, (2) the latent masked image $Z(I_p)$, and (3) a mask I_{m}, forming a 9-channel input. The U-Net's initial convolution is expanded accordingly, with new weights initialized from the pretrained network. The clothing image I_{c} is processed by the CLIP encoder for exemplar-based conditioning. During training, inputs $\{Z(I_{pt}), Z(I_{pm}), I_m\}$ are concatenated and fed into the network. The Diffusion U-Net predicts a denoised version of the input, leveraging the global condition $CLIP(I_c)$ integrated via cross-attention. The loss function is defined as:

$$L_{\text{simple}} = \left|\left|\epsilon - \epsilon_\theta\left(Z(I_{pt}), Z(I_{pm}), I_m, \text{CLIP}(I_c), t\right)\right|\right|^2. \quad (1)$$

This approach effectively combines person and clothing representations, enhancing the denoising process with contextual cross-attention for improved virtual try-on outputs.

E. Additional Feature Preserving Block (ControlNet)

The Additional Feature Preserving Block (Training Loop 2, Fig. 2 integrates external conditions into the architecture to enhance neural network performance. Inspired by ControlNet [26], it introduces conditioning inputs like edge maps or segmentation masks to enhance image generation control without modifying the original model. Using pre-trained blocks, it employs two spatial encoders: one extracts garment-specific features from the clothing image I_c, while the other processes pose representation I_{openpose}. Both use a U-Net architecture, with the pose encoder integrating intermediate features into the main U-Net via the Additional Feature Preserving Block for alignment and realism. Training uses paired clothing and image datasets, compensating for the lack of paired data of the same person in different outfits and identical poses.

F. Experimental setting

1) Datasets

Our experiments primarily utilize the VITON-HD dataset [22], which includes 13,679 pairs of frontal-view images of women and top clothing at a resolution of 1024×768. Consistent with previous studies [6, 10, 22], we divide the dataset into a training set with 11,647 pairs and a test set with 2,032 pairs, conducting experiments at 1024×768 resolution.

TABLE I. TTRAINING STEP PROCESS

Algorithm Training
1. repeat
2. random t with $t \sim \text{Uniform}(\{1, \dots, T\})$
3. random ϵ (noise) with $\epsilon \sim \mathcal{N}(0, I)$
4. make $Z(I_p)$ noisy image with the noise from 3 and timestep from 2
5. Take gradient descent step on $L_{\text{simple}} = \left|\left|\epsilon - \epsilon_\theta\left(Z(I_{pt}), Z(I_{pm}), I_m, \text{CLIP}(I_c), t\right)\right|\right|^2$
6. until converged

2) Evaluation Metrics

We use several common metrics to assess our method's performance under two test settings. The first setting is the paired cloth and model setting, where the clothes image is used to reconstruct the person image. In this setting, we utilize LPIPS (Learned Perceptual Image Patch Similarity) and SSIM (Structural Similarity Index), which assess perceptual and structural similarity, respectively. The second setting is the unpaired setting, where FID (Fréchet Inception Distance) and KID (Kernel Inception Distance) are used to evaluate the quality and diversity of the generated images. These metrics provide a comprehensive evaluation of both perceptual accuracy and image quality in virtual try-on scenarios.

3) Implementation Details

We train the two stages of our framework separately. For the first diffusion process, we employ a latent space autoencoder with a down sampling factor of $f = 8$, meaning the latent space's spatial dimensions are $c \times \left(\frac{H}{f}\right) \times \left(\frac{W}{f}\right)$ with a channel dimension $c = 4$ The denoising UNET is based on [27], utilizing the $AdamW$ optimizer with a learning rate of 1×10^e $- 7$. We initialize the model using the setup from [27], allowing it to learn basic inpainting, and then continue training for 50 epochs on NVIDIA Tesla A5000 GPUs. For inference, we use the DDPM method [4] with 50 sampling steps. For the second stage of our framework, which trains the semantic encoder (ControlNet), the U-Net down block is based on [27], with an Additional Feature Preserving Block added at the end before injecting the residuals. The AdamW optimizer is used with a learning rate of 1×10^e –5, and the following hyperparameters are configured, Condition Embedding Channel as [384, 768, 1152, 1536], conditioning scale as 0.8, and guidance scale as 5. During inference, we adopt the DDPM method [4] with 50 sampling steps. We continue training from stage one of our framework for an additional 20 epochs on NVIDIA Tesla A5000 GPUs. For inference, we again use the DDPM method [4] with 50 sampling steps. Result and Discussion

G. Quantitative Evaluation

From Table 2. our method outperforms the compared models across all evaluation metrics. Specifically, LPIPS [30] and SSIM [23] are evaluated in a paired setting with paired images and paired clothes, while KID [25] and FID [24] are assessed in an unpaired setting, demonstrating significant improvements in image quality and perceptual similarity.

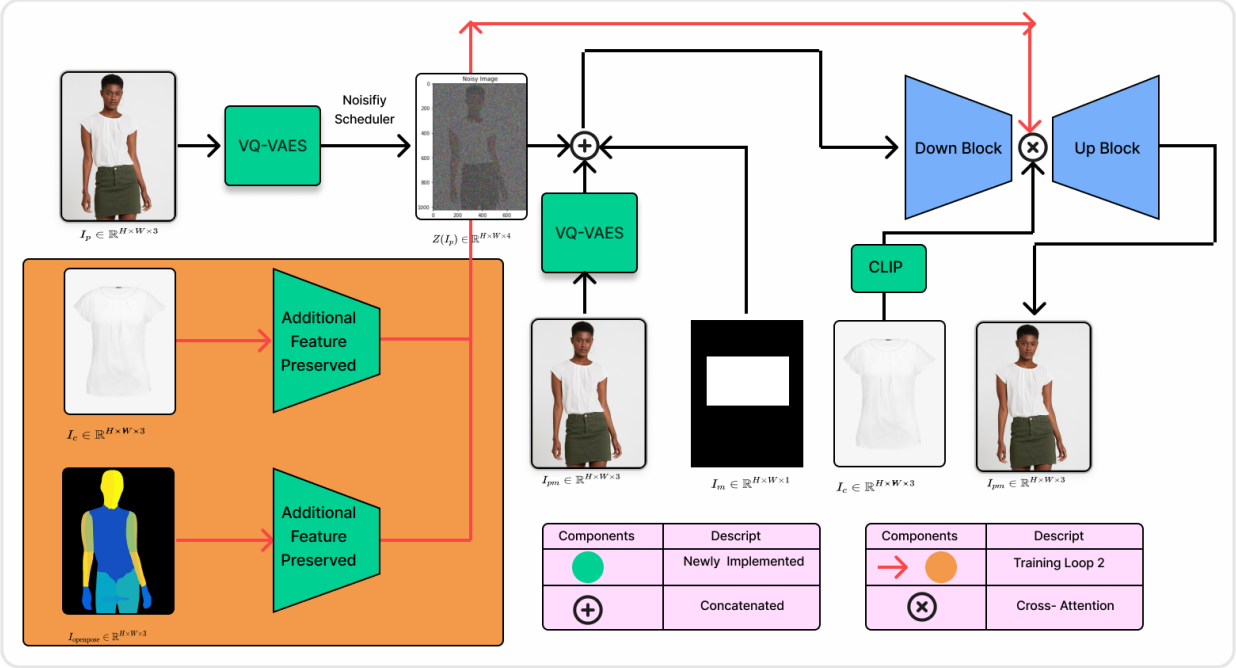

Fig. 2. In Training Loop 1 (excluding orange box), the U-Net is trained with a diffusion model on inpainting tasks to capture a coarse synthesis of the image, effectively learning both global and local features. In Training Loop 2 (Orange Box), the U-Net from Training Loop 1 is frozen, and the Additional Feature Preserving Block is trained to add residual details. This step enhances fine-grained image features by incorporating additional cross-attention guidance, refining the overall output

With the lowest LPIPS score (0.0820), it surpasses state-of-the-art models such as VITON-HD (0.117) and HR-VTON (0.097), indicating that our approach produces more perceptually accurate results. Additionally, our method achieves the best FID (7.782) and KID (1.53), suggesting that it generates images closer to real images in terms of feature distribution and diversity. While SSIM is slightly lower (0.825) compared to LADI-VTON (0.864) and HR-VTON (0.878), this reflects the trade-off between maintaining structural similarity and capturing more intricate details and textures, which is crucial for realistic virtual try-on applications. Overall, our method's performance highlights the robustness of the diffusion model and its ability to generate high-quality, realistic virtual try-on images.

H. Qualitative Evaluation

We conducted a comprehensive evaluation of our TryOn Framework against state-of-the-art methods, including LADI-VTON [10], VITON-HD, and HR-VTON [22], using their publicly available implementations. For a fair comparison, we focused on models trained on the VITON-HD dataset at a resolution of 1024×768. Fig. 3 and 4 highlight the superior performance of our TryOn Framework in generating high-quality 1024×768 images, excelling at preserving intricate clothing details such as logos and textures, thanks to its innovative Additional Feature Preserving Block. Additionally, it synthesizes natural and realistic body shapes regardless of the clothing in the reference image. In contrast, GAN-based methods like VITON-HD [6] and HR-VTON [22]-

TABLE II. QUANTITATIVE COMPARISONS IN SINGLE DATASET SETTINGS, VITON-HD DATASETS.

Method	LPIPS	FID	KID	SSIM
VITON-HD [6]	0.117	12.117	3.23	0.862
LADI-VTON [10]	0.096	9.480	1.99	0.864
HR-VTON [22]	0.097	11.265	2.73	0.878
OursTryon	0.082 ↑	7.782 ↑	1.53 ↑	0.825↓

-retain logos and text but show significant artifacts, especially in handling posture variations like crossed arms, which compromise realism. Similarly, while LADI-VTON [10] uses diffusion-based approaches to reduce artifacts and produce cleaner images, it struggles to preserve original facial features, accurately render fabric textures, and maintain fine clothing details, as shown in Fig. 3 and 4, and produces images at a lower resolution. By overcoming these challenges, our TryOn Framework accurately reproduces facial features, generates realistic textures, and preserves fine clothing details, making it a significant improvement over both GAN-based and diffusion-based methods and consistently delivering lifelike, visually superior results.

IV. CONCLUSION

Through the integration of advanced techniques, our approach significantly enhances the realism, accuracy, and practicality of virtual try-on systems, setting a new benchmark for future developments. Our first contribution is the integration of Vector Quantized Variational Autoencoders (VQ-VAEs) into the virtual try-on framework, marking their first use in this context. This integration enhances compatibility with the

diffusion U- Net architecture, improving feature matching and image synthesis, ultimately resulting in more realistic outputs.

Second, we shift the conditioning process from text- based to image-based input by using the CLIP image encoder instead of the CLIP text encoder employed in methods like LADI-VTON. This enables the system to directly leverage visual information from clothing images, eliminating the need for descriptive text. By focusing on visual conditioning, we achieve a more accurate and realistic representation of garments in generated try-on images, aligning better with the needs of virtual try-on systems.

Finally, we address a key challenge in virtual try-on—preserving fine-grained clothing details such as logos and textures—by -introducing an Additional Feature Preserving Block within the diffusion model. These blocks retain intricate garment features during the generation process, ensuring higher fidelity in the outputs. This novel approach has not been previously applied in virtual try-on systems.

Looking ahead, future research could explore multimodal conditioning by incorporating both visual and attribute-based inputs, such as garment dimensions and model characteristics. Additionally, introducing perceptual loss during the training process could further enhance the visual quality of the generated images. Improving performance in paired settings, where spatial fidelity is critical, also remains an important area for future work. Overall, our framework provides a strong foundation for advancing virtual try-on systems, combining cutting-edge techniques to enhance both realism and quality.

Fig. 3. Qualitative results generated by Our TryOn architecture and competitors in paired setting

Fig. 4. Qualitative results generated by Our TryOn architecture and competitors in unpaired setting

V. REFERENCES

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

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

[3] Dhariwal, P. and Nichol, A., 2021. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34, pp.8780-8794

[4] Ho, J., Jain, A. and Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, pp.6840-6851.

[5] He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[6] Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587)

[7] Yu, R., Wang, X. and Xie, X., 2019. Vtnfp: An image-based virtual try-on network with body andclothing feature preservation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10511-10520).

[8] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. and Ganguli, S., 2015, June. Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning (pp. 2256-2265). PMLR.

[9] Cucurull, G., Taslakian, P. and Vazquez, D., 2019. Context-aware visual compatibility prediction. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp.12617- 12626).

[10] Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M. and Cucchiara, R., 2023. LaDI-VTON:Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. arXiv preprint arXiv:2305.13501.

[11] Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W. and Luo, P., 2020.Towards photo-realistic virtual tryon by adaptively generating-preserving image content. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7850- 7859).

[12] Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet,D. and Norouzi, M., 2022, July. Palette: Image-to-image diffusion models. In ACM SIGGRAPH 2022 Conference Proceedings (pp. 1- 10).

[13] Issenhuth, T., Mary, J. and Calauzenes, C., 2020. Do not mask what you do not need to mask: a parser-free virtual try-on. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16 (pp. 619-635). Springer International Publishing.

[14] Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J. and Norouzi, M., 2022. Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4), pp.4713-4726.

[15] Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M. and Lin, L., 2019.Graphonomy: Universal human parsing via graph transfer learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7450- 7459).

[16] Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C. and Murphy, K., 2017. Towards accurate multi-person pose estimation in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4903-4911).

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

[18] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision.In *International conference on machine learning* (pp. 8748-8763). PMLR.

[19] Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi,M. and Kemelmacher-Shlizerman, I., 2023. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4606-4615).

[20] Rocco, I., Arandjelovic, R. and Sivic, J., 2017. Convolutional neural network architecture for geometric matching. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6148-6157).

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

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

[23] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. TIP (2004).

[24] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, *30*.

[25] Bińkowski, M., Sutherland, D.J., Arbel, M. and Gretton, A., 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

[26] Zhang, L., Rao, A. and Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3836-3847).

[27] Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A. and Dimitrov, D., 2023. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*.

[28] Van Den Oord, A. and Vinyals, O., 2017. Neural discrete representation learning. *Advances in neural information processing systems*, *30*.

[29] He, S., Song, Y.Z. and Xiang, T., 2022. Style-based global appearance flow for virtual try-on. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3470-3479).

[30] Zhang, R., Isola, P., Efros, A.A., Shechtman, E. and Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 586-595).