Meeting Note

Author: Henry Lin
Date: 2024/08/12 – 2024/08/16

Weekly Summary
IMAGDressing-v1 [1].

Plan for Next Week
U-Net (*arXiv*: 1505.04597)

Details
The concept of virtual try-on (VTON) has been proposed by Han et al. [2] since 2018, but the control over optional faces, poses, and scenes are still neglected. The early VTON typically use the generative adversarial network (GAN) and two-stage strategy, they first warp the clothing into a specifical shape, then use GAN to combine the masked human image with the clothing image. In recent year, researches start to use diffusion model to replace GAN, and in 2023, Zhu et al. [3] introduced a parallel U-Nets architecture.

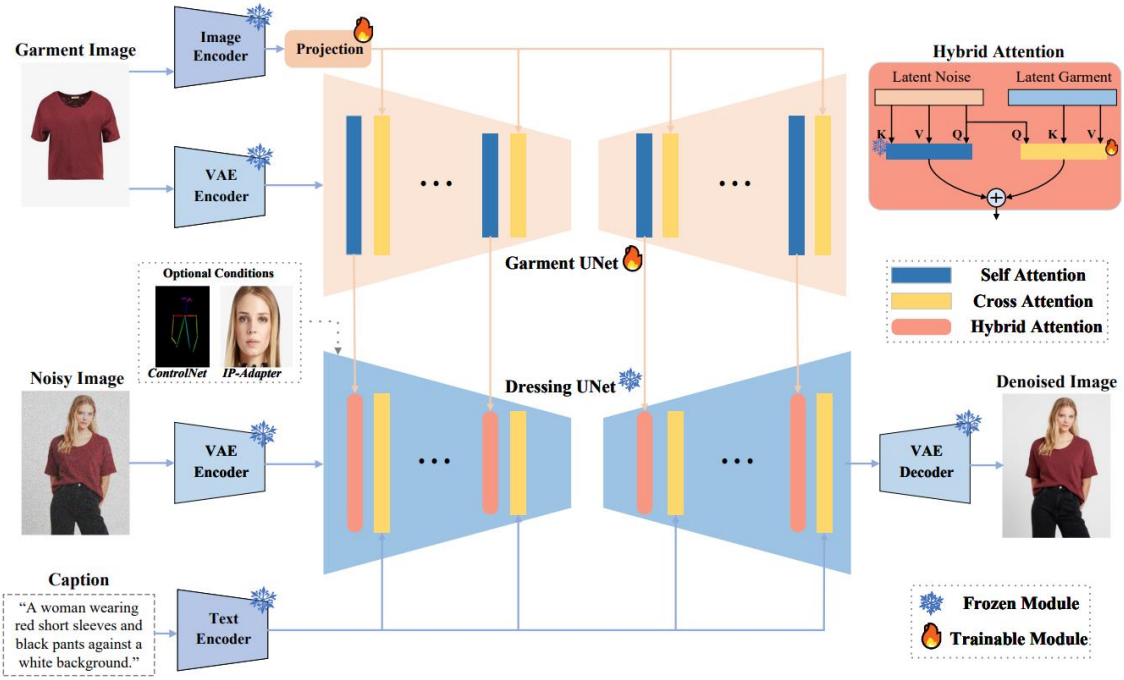

Figure 1. The architecture of proposed IMAGDressing-v1.

To address the problem of flexibility, the authors proposed an architecture, which is compatible with other community model such as ControlNet [7] and IP-Adapter [8].

Figure 2. IMAGDressing-v1 under specific conditions.

The garment U-Net is a Stable Diffusion V1.5 (SD v1.5) [4] model, used to extract the fine-grained garment features. The input is obtained from a pre-trained frozen VAE encoder, which is used to extract the garment semantic features; and the condition input used to extract the texture features, is obtained from a pre-trained frozen CLIP image encoder [5] and a trainable Q-Former [6] based projector. As the Figure 1 shows, the garment semantic features and the texture features are interacting though the cross-attention mechanism.

The dressing U-Net is also a SD v1.5 based model, but all self-attention layers are replaced with hybrid attention modules to integrate the garment features. The output Z_h of hybrid attention is as follows:

$$\mathbf{Z}_h = \underbrace{\text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}}_{\text{Self Attention}} + \lambda \underbrace{\text{Softmax}\left(\frac{\mathbf{Q}\left(\mathbf{K}'\right)^\top}{\sqrt{d}}\right)V'}_{\text{Cross Attention}}$$

inside the cross-attention of hybrid attention module, query Q is shared, and the key K' and value V' is obtained from garment U-Net. The weight of self-attention is frozen and initialed using the self-attention layers in pre-trained SD

v1.5 model, while the cross-attention is trainable, this architecture has two major objectives: (1) maintaining the original generation abilities, and (2) incorporating additional garment features. And for the text condition, the features are extracted from the pre-trained frozen CLIP text encoder, and enabling the model to control the scenes through text.

The loss function is similar to the original loss function of latent diffusion model (LDM) as follows:

$$L_{LDM} = \mathbb{E}_{\mathbf{z}_t, \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \mathbf{C}_t, \mathbf{C}_g, t} \left\| \epsilon_\theta \left(\mathbf{z}_t, \mathbf{C}_t, \mathbf{C}_g, t \right) - \epsilon_t \right\|^2$$

where ϵ_θ denotes the denoising U-Net (i.e. dressing U-Net), $z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t} - \epsilon_t$, $z_0 = \varepsilon(x_0)$, and ε, x_0, C_t, C_g, t represent VAE encoder, real image, text condition, garment features, and timestamp, respectively.

Remarks

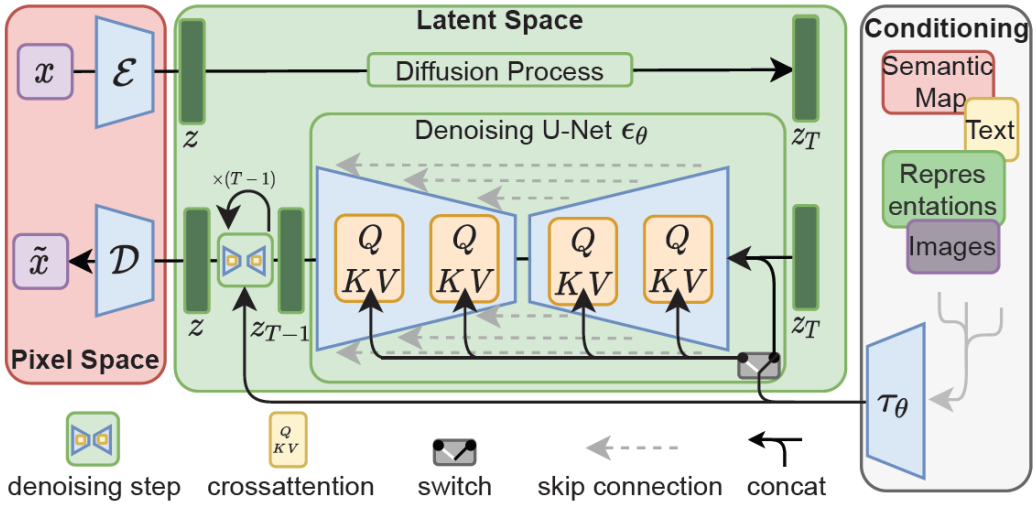

Figure 3. The architecture of LDM.

In LDM, the conditional input (e.g. text and images, etc.) is integrated into the denoising U-Net by connecting to the key and value fields of the cross-attention modules.

In this paper, the authors didn't clearly describe how actually is it to combine

IMAGDressing-v1 with ControlNet or other modules.

Table 1. Comparison of IMAGDressing-v1 and other diffusion models.

Method	ImageReward (↑)	MP-LPIPS (↓)	CAMI-U (↑)	CAMI-S (↑)
Blip-Diffusion	-2.224	0.1824	1.051	-
Versatile Diffusion	-2.055	0.4321	1.253	-
IP-Adapter	-2.267	0.4093	1.381	-
MagicClothing	-0.164	0.1499	1.655	2.692
Ours	**-0.095**	**0.1466**	**1.753**	**2.719**

The authors compared the proposed model with other diffusion models. As the authors said, these models are not specifically designed for VTON tasks except MagicClothing, and even only the MP-LPIPS [9] metric is designed for virtual try-on tasks.

References

[1] Shen, F., Jiang, X., He, X., Ye, H., Wang, C., Du, X., ... & Tang, J. (2024). IMAGDressing-v1: Customizable Virtual Dressing. *arXiv preprint arXiv:2407.12705*.

[2] HAN, Xintong, et al. Viton: An image-based virtual try-on network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. p. 7543-7552.

[3] ZHU, Luyang, et al. Tryondiffusion: A tale of two unets. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 4606-4615.

[4] ROMBACH, Robin, et al. High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022. p. 10684-10695.

[5] RADFORD, Alec, et al. Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. PMLR, 2021. p. 8748-8763.

[6] LI, Junnan, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *International conference on machine learning*. PMLR, 2023. p. 19730-19742.

[7] ZHANG, Lvmin; RAO, Anyi; AGRAWALA, Maneesh. Adding conditional control to text-to-image diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023. p. 3836-3847.

[8] YE, Hu, et al. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.
[9] CHEN, Weifeng, et al. Magic Clothing: Controllable Garment-Driven Image Synthesis. *arXiv preprint arXiv:2404.09512*, 2024.