ICCV #11745

Learning Implicit Features with Flow Infused Attention for Realistic Virtual Try-On

Supplementary Material

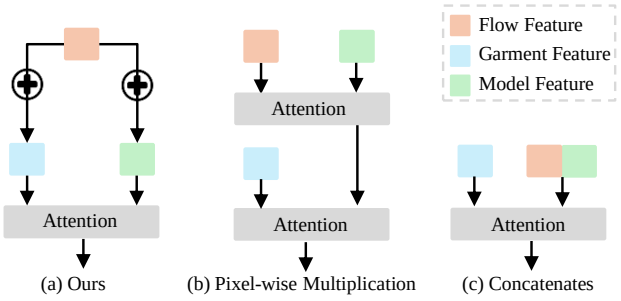

Figure S1. Detailed design of variants of FIA. All output features are fused with the spatial features through cross-attention. For simplicity, this process is omitted in the figure.

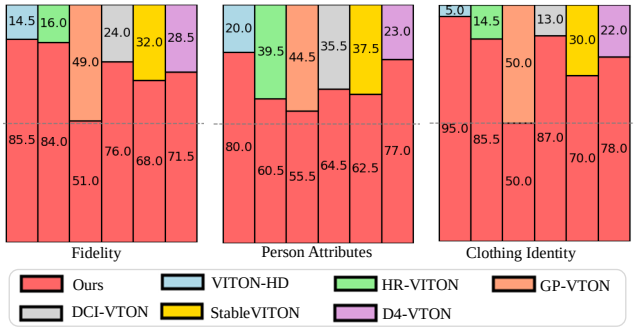

Figure S2. User study results on VITON-HD dataset. We compare our FIA-VTON with six baselines, involving a total of 40 participants.

A. Implementation details

We use the VAE and denoising U-Net from Stable Diffusion v2.1, initializing with the weights from SD 2.1 (using zero initialization if the corresponding structure is matched). The GarmentNet and the Denoising U-Net have similar architectures, but the GarmentNet does not incorporate flow-infused attention. The GarmentNet and U-Net each output 21 feature maps, with 9 from the encoder and 12 from the decoder. The 21 output feature maps from GarmentNet are injected into the denoising U-Net via flow-infused attention, according to their output sizes and order.

For the Flow Guider, we adopt the Dynamic Semantics Disentangling Module (DSDM) from $\mathcal{D}^4$-VTON [7] to estimate a dense warp flow at the same resolution as the original garment image. Taking a garment with a resolution of 512 $\times$ 384 as an example, the Flow Guider generates a flow of the same 512 $\times$ 384 resolution, while GarmentNet/U-Net generates feature maps with the following resolutions: 64 $\times$ 48, 32 $\times$ 24, 16 $\times$ 12, and 8 $\times$ 6. To correctly implement Equation 1 from the main paper, we downsample the flow to different sizes using bilinear interpolation and add it to the feature map of the corresponding size.

B. User Study

We conduct a user study with 40 participants for models trained on the VITON-HD dataset at 512 $\times$ 384 resolution. Each participant compared an image from the baseline with one from our model based on the following criterion:

- **Fidelity**: Select the image that better reflects realism in terms of human body and colour harmony
- **Person Attributes**: Select the image that better preserves the skin colour, pose and appearance features of the person.
- **Clothing Identities**: Select images that better preserve the design, texture details, logos and shapes of the clothing.

As shown in Figure S2, FIA-VTON outperforms other methods across all criteria, especially in clothing identity, demonstrating its excellent ability to preserve clothing texture.

C. In-the-wild senarios

Furthermore, we evaluate the wild scenarios to test the robustness and applicability of FIA-VTON in real-world conditions. As shown in Figure S3, FIA-VTON accurately recognizes and integrates the shape of complex garments, such as off-shoulder designs, with the person. It can generate interlaced parts for complex poses, such as sitting. Additionally, it effectively completes and integrates the background with the garment in complex in-the-wild scenarios.

D. High Resolution Results

In order to show the suitability of our method for high-resolution results, we train FIA-VTON on VITON-HD [1] at 1024 $\times$ 768 resolution. Instead of training from scratch, we progressively train 30,000 iterations from the previously trained weights and test it on the corresponding 1024 $\times$ 768 resolution test set.

Qualitative Results. In Figure S4, we demonstrate the qualitative results of FIA-VTON on the VITON-HD dataset at 1024 $\times$ 768 resolution. We randomly select 6 persons and 6 garments for virtual try-on. The generated results

Figure S3. Qualitative comparison in the wild scenarios. Compared with state-of-the-art methods (OOTDiffusion[6], CAT-VTON[3]). Our method generates more natural images that seamlessly combine background, person, and garment in complex scenarios. Zoom in for more details.

show that our model can handle style changes well and retain the garment details effectively for different models and garments.

Method Comparison. We select methods (HR-VITON [4], GP-VTON [5], IDM-VTON [2], CAT-VTON [3]) that support 1024 × 768 resolution as a comparison with our method. As Figure S5 shown, our FIA-VTON demonstrates a superior ability to retain fine-grained texture details while preserving garment-specific features such as seams, logos, and complex fabric patterns.

E. Disuccsion

Limitation. Our FIA-VTON requires additional conditional inputs to support the warp model, contributing to increased inference time. Explore integrating the warp model directly with the diffusion process during training, allowing for the possibility of discarding the warp model entirely during inference, is left as future work.

Potential negative impact. While this work proposes a diffusion-based method that offers significant benefits in generating realistic and personalized virtual try-on experiences, it also comes with potential concerns. The method could be misused for creating deceptive or non-consensual imagery, potentially harming individuals' privacy or misrepresenting their appearance. Therefore, users and developers must apply this technology responsibly, ensuring that usage respects privacy, ownership, and ethical standards to prevent any form of malicious exploitation.

References

[1] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul 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, pages 14131–14140, 2021. 1

[2] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on. arXiv preprint arXiv:2403.05139, 2024. 2

[3] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models, 2024. 2

[4] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In European Conference on Computer Vision, pages 204–219. Springer, 2022. 2

[5] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. 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, pages 23550–23559, 2023. 2

[6] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. arXiv preprint arXiv:2403.01779, 2024. 2

[7] Zhaotong Yang, Zicheng Jiang, Xinzhe Li, Huiyu Zhou, Junyu Dong, Huaidong Zhang, and Yong Du. D4-vton: Dynamic semantics disentangling for differential diffusion based virtual try-on. arXiv preprint arXiv:2407.15111, 2024. 1

ICCV
#11745

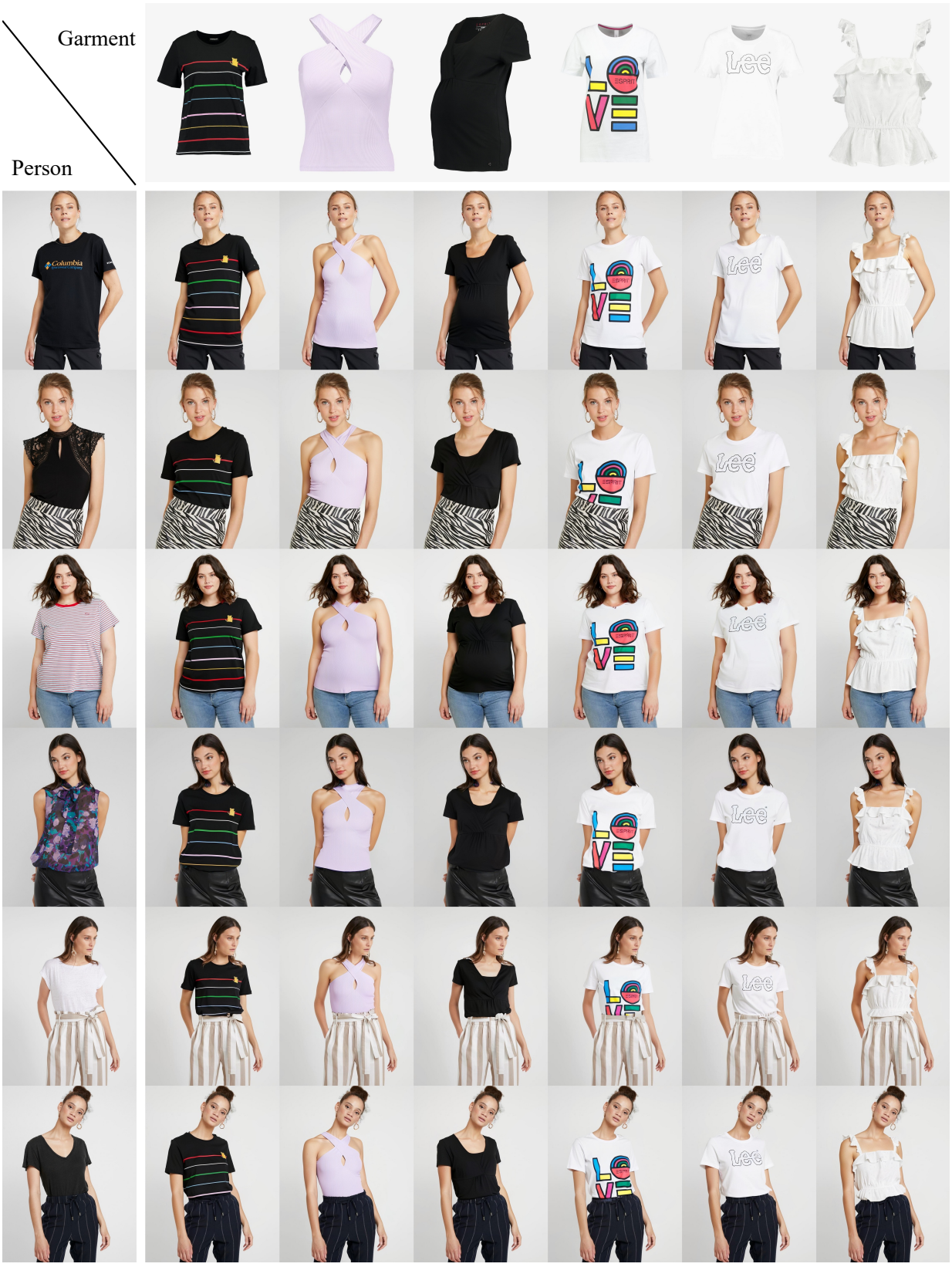

Figure S4. Try-on results on VITON-HD test data by FIA-VTON trained on VITON-HD training data at 1024 × 768 resolution. Best viewed in zoomed, color monitor.

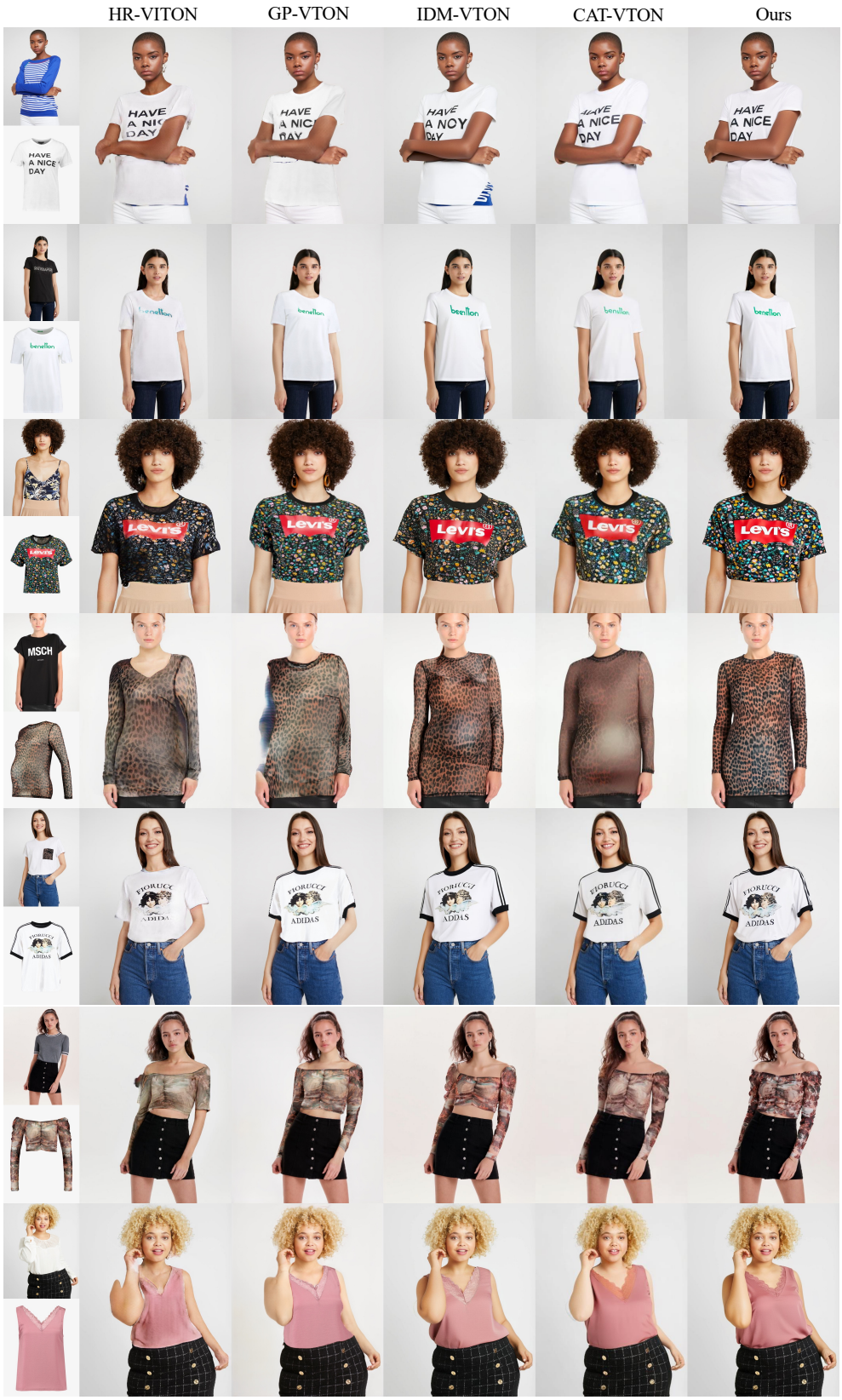

Figure S5. Qualitative comparison on VITON-HD dataset at 1024 × 768 resolution. Best viewed in zoomed, color monitor.