Mitigating Occlusions in Virtual Try-On via A Simple-Yet-Effective Mask-Free Framework

Chenghu Du[1] **Shengwu Xiong**[3] **Junyin Wang**[1]

Yi Rong[1,4,*] **Shili Xiong**[1,3,*]

[1]Wuhan University of Technology, [2]Shanghai AI Laboratory

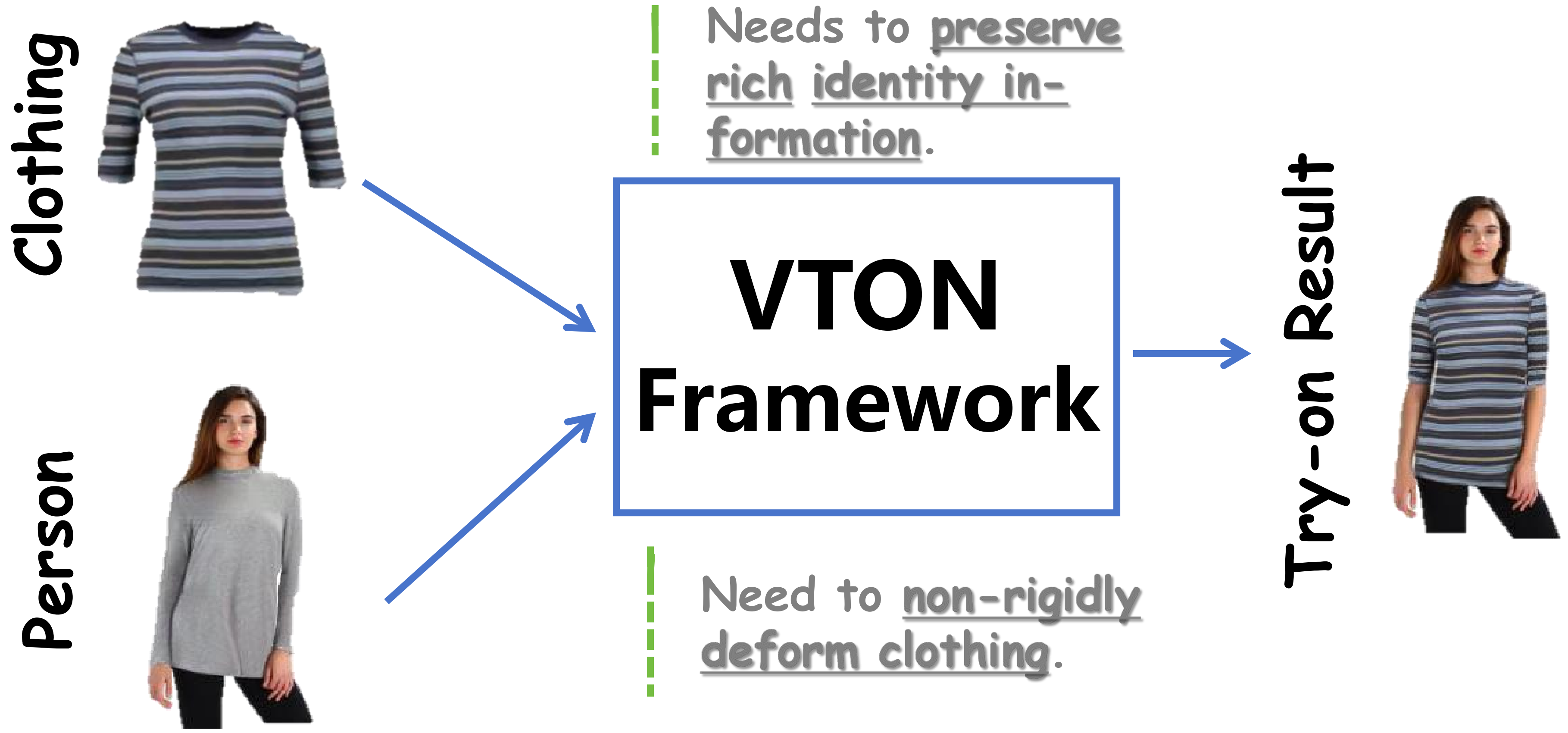
[3]Interdisciplinary Artificial Intelligence Research Institute, Wuhan College

[4]Sanya Science and Education Innovation Park, Wuhan University of Technology

{duch, xiongsw, wjy199708, yrong}@whut.edu.cn

Introduction

NEURAL INFORMATION PROCESSING SYSTEMS



This Work aims at transfer a target clothing onto a specific person.

Motivations

Visualization of the occlusion issues in VTON and the effectiveness of the proposed method.} It illustrates the inherent occlusion (Cyan regions) caused by imprecise inpainting masks and acquired occlusion (Red regions) resulting from erroneous human structural representations.

Motivations

- Virtual Try-On (VTON) suffers from two critical occlusion problems that degrade generation quality:
- 🎯 Inherent Occlusions (Cyan regions in Fig.1)
- *Cause*: Imprecise inpainting masks leave residual clothing areas in input images
- *Effect*: Model mistakenly preserves clothing "ghosts" from reference images
- *Impact*: Sub-optimal associations between target clothing and background/body pixels
- 🎯 Acquired Occlusions (Red regions in Fig.1)
- *Cause*: Erroneous Human Structural Representations (HSRs) misguide generation
- *Effect*: Disrupted spatial structures of body parts (missing limbs, distorted anatomy)
- *Impact*: Unreasonable human body generation, especially with varying clothing coverage
- ***Key Insight***: Both mask-based and mask-free methods are vulnerable due to reliance on imperfect masks and HSRs

Proposed Framework

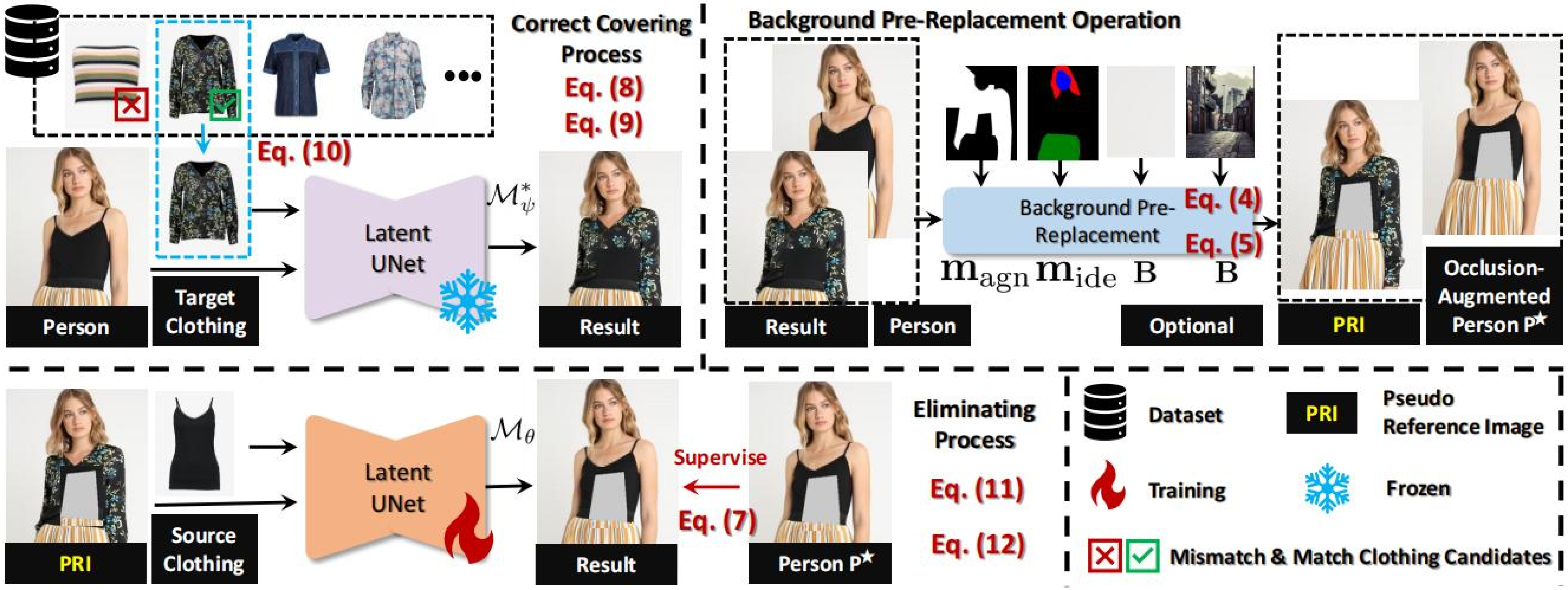


Figure 2: **Overview of our proposed framework.** It includes two operations: the covering and eliminating processes to integrate clothing with a person image, and the background pre-replacement operation to create a pseudo reference image by substituting the original background.

Proposed Framework

- No masks or HSRs required at inference
- Plug-and-play compatibility with GANs & Diffusion Models
- In-the-wild generalization capability

- 🔧**Key Operation 1**: Background Pre-Replacement (For Inherent Occlusion)
- Mechanism: Replace background regions before training to sever clothing-background associations.

- 🔧**Key Operation 2**: Covering-and-Eliminating (For Acquired Occlusion)
- Mechanism: Train model to eliminate larger covering clothing and reconstruct underlying body.

Experiments

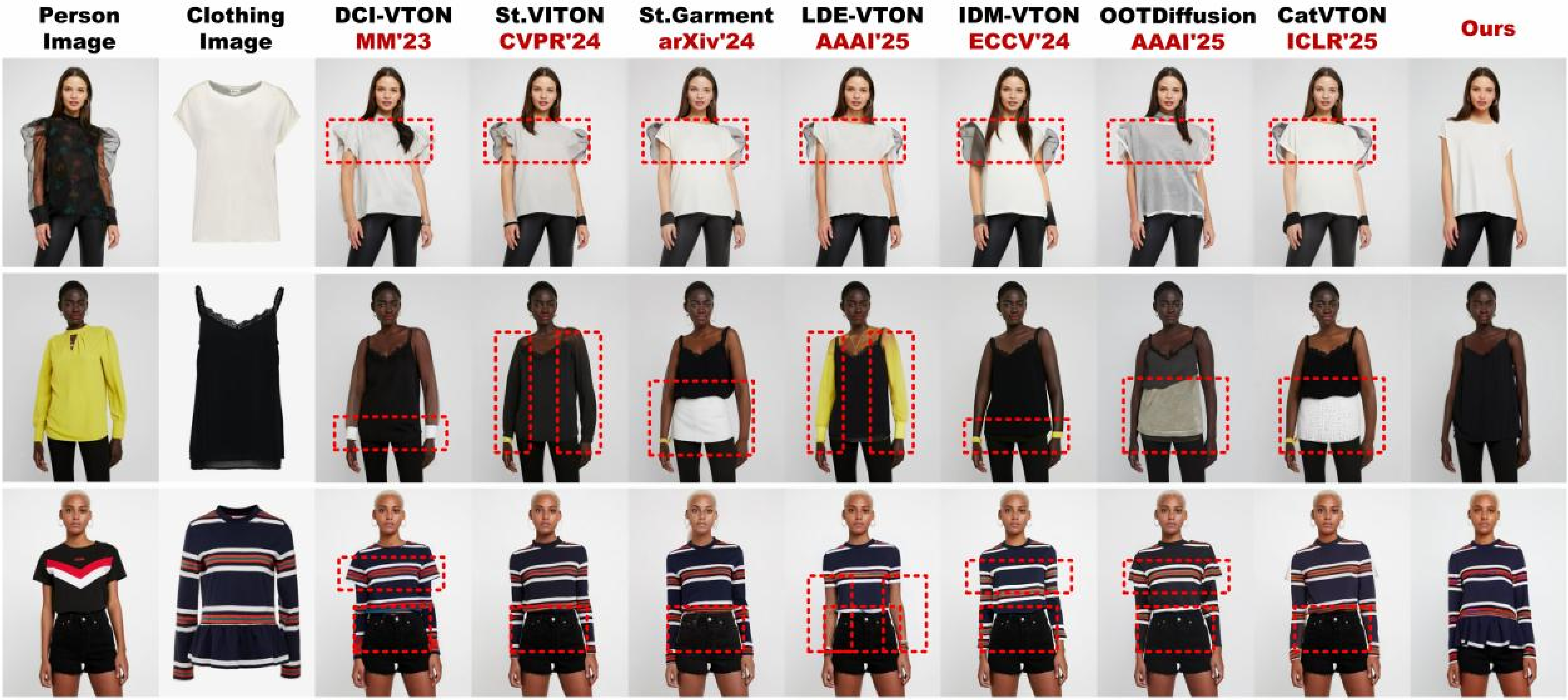


Figure 3: **Qualitative results** on the VITON-HD dataset. The baseline methods consist of seven SOTA diffusion-based methods. Red dashed boxes highlight the limitations of each method.

Experiments

Figure 4: **Qualitative results** on the DressCode dataset. The baseline methods consist of four SOTA diffusion-based methods. Red dashed boxes highlight the limitations of each method.

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

Table 2: **Quantitative comparisons on the VITON-HD and DressCode datasets.** For LPIPS, FID, and KID, the lower the better. For SSIM, the higher the better."Mask-Free" denotes whether the inpainting mask $\mathbf{m}_{\text{agn}}$ and human structural representation (HSR) $\mathbf{r}$ are used during *inference*. **Bold** denotes the best result. $\underline{\text{Underline}}$ represents second best.

Train / Test				VITON-HD				DressCode Upper			
Methods	Publication	Backbone	Mask-Free	$\text{SSIM}_p \uparrow$	$\text{LPIPS}_p \downarrow$	$\text{FID}_{up} \downarrow$	$\text{KID}_{up} \downarrow$	$\text{SSIM}_p \uparrow$	$\text{LPIPS}_p \downarrow$	$\text{FID}_{up} \downarrow$	$\text{KID}_{up} \downarrow$
VITON-HD [21]	CVPR'21	ResUnet	✗	0.862	0.117	12.117	3.23	n/a	n/a	n/a	n/a
HR-VITON [26]	ECCV'22	ResUnet	✗	0.878	0.105	11.265	2.73	$\underline{0.936}$	0.065	13.82	2.71
GP-VTON [4]	CVPR'23	ResUnet	✗	0.884	0.081	9.701	1.26	0.769	0.270	20.11	8.17
LaDI-VTON [10]	MM'23	SD1.5	✗	0.864	0.096	9.480	1.99	0.915	0.063	14.26	3.33
PbE [11]	CVPR'23	SD1.5	✗	0.802	0.143	11.939	3.85	0.897	0.078	15.33	4.64
DCI-VTON [5]	MM'23	SD1.5	✗	0.880	0.080	8.998	1.19	**0.937**	**0.042**	11.92	1.89
StableVITON [7]	CVPR'24	SD1.5	✗	0.864	0.084	9.465	1.40	n/a	n/a	n/a	n/a
StableGarment [27]	arXiv'24	SD1.5	✗	0.803	0.104	17.115	8.85	n/a	n/a	n/a	n/a
Anydoor [13]	CVPR'24	SD1.5	✗	0.821	0.099	10.850	2.46	0.899	0.119	14.83	3.05
IDM-VTON [28]	ECCV'24	SDXL	✗	0.850	$\underline{0.060}$	9.842	1.12	0.880	0.056	9.54	4.32
LDE-VTON [16]	AAAI'25	SD1.5	✗	0.884	0.081	9.640	1.21	n/a	n/a	n/a	n/a
CatVTON [9]	ICLR'25	SD1.5	✗	0.870	0.061	9.287	1.17	0.902	$\underline{0.045}$	$\underline{7.40}$	2.62
BooW-VTON [14]	CVPR'25	SDXL	✓	0.862	0.108	**8.809**	**0.82**	0.919	0.062	11.03	**0.86**
Ours	This Work	SD1.5	✓	**0.889**	**0.057**	$\underline{8.854}$	$\underline{0.96}$	0.923	**0.042**	**6.58**	$\underline{1.72}$

- n/a: official code or data is inaccessible.

Conclusion

- ☑ Novel Framework: First unified mask-free VTON framework addressing both occlusion types simultaneously
- ☑ Background Pre-Replacement: Simple operation that fundamentally eliminates inherent occlusion by preventing clothing-background confusion
- ☑ Covering-and-Eliminating: Training strategy that enhances human structure understanding and mitigates acquired occlusion without explicit HSRs
- ☑ Plug-and-Play Design: Compatible with any generative architecture (validated on GANs & Diffusion Models)
- ☑ In-the-Wild Ready: Robust to diverse backgrounds via synthetic background augmentation
- ☑ SOTA Performance: Outperforms 21 recent methods on 3 datasets while being truly mask-free

Thank You!