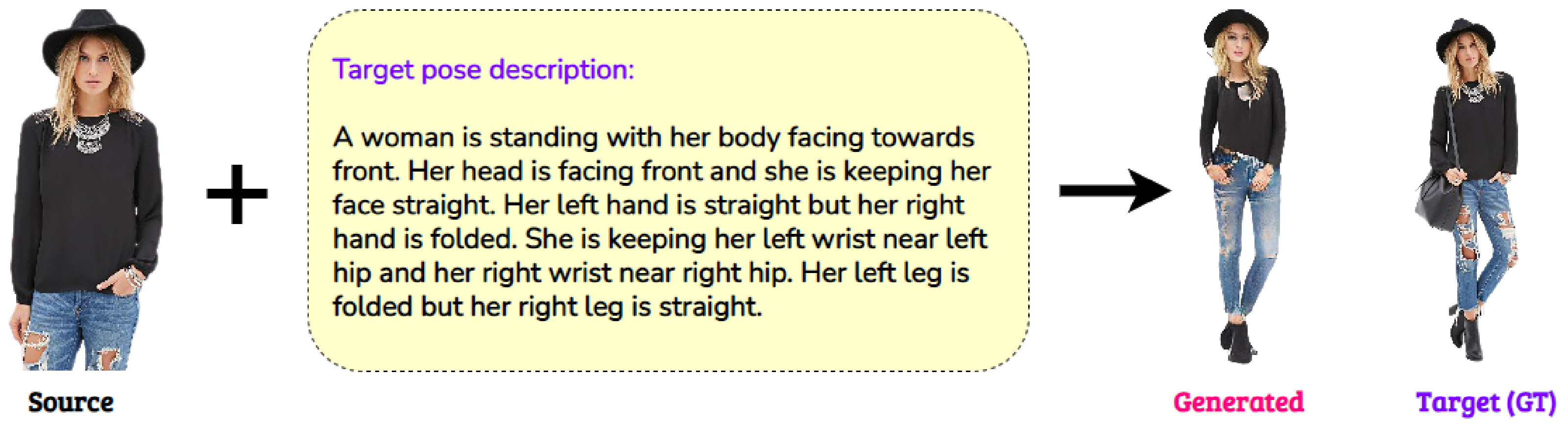
TIPS: Text-Induced Pose Synthesis

P. Roy[1] S. Ghosh[1] S. Bhattacharya[2] U. Pal[3] M. Blumenstein[1]

[1]University of Technology Sydney [2]IIT Kharagpur [3]ISI Kolkata

ECCV TEL AVIV 2022

Introduction



Target pose description:

A woman is standing with her body facing towards front. Her head is facing front and she is keeping her face straight. Her left hand is straight but her right hand is folded. She is keeping her left wrist near left hip and her right wrist near right hip. Her left leg is folded but her right leg is straight.

We introduce a novel text-supervised human pose synthesis technique to mitigate the structural inconsistencies in traditional keypoints-guided pose transfer schemes.

Main Contributions

- A novel three-stage sequential pipeline for text-guided human pose synthesis.
- A new dataset DF-PASS by extending the DeepFashion dataset with human-annotated text descriptions of poses.

Analytical Results

Table 1. Performance of pose transfer algorithms on DeepFashion.

Pose Generation Algorithm	SSIM	IS	DS	PCKh	GCR	LPIPS (VGG)	LPIPS (SqzNet)
Partially Text Guided (Ours)	0.549	3.269	0.950	0.53	0.963	0.402	0.290
Fully Text Guided (Ours)	0.549	3.296	0.950	0.53	0.963	0.402	0.289
Zhou et al.	0.373	2.320	0.864	0.62	0.979	0.310	0.215
PATN	0.773	3.209	0.976	0.96	0.983	0.299	0.170
Real Data	1.000	3.790	0.948	1.00	0.995	0.000	0.000

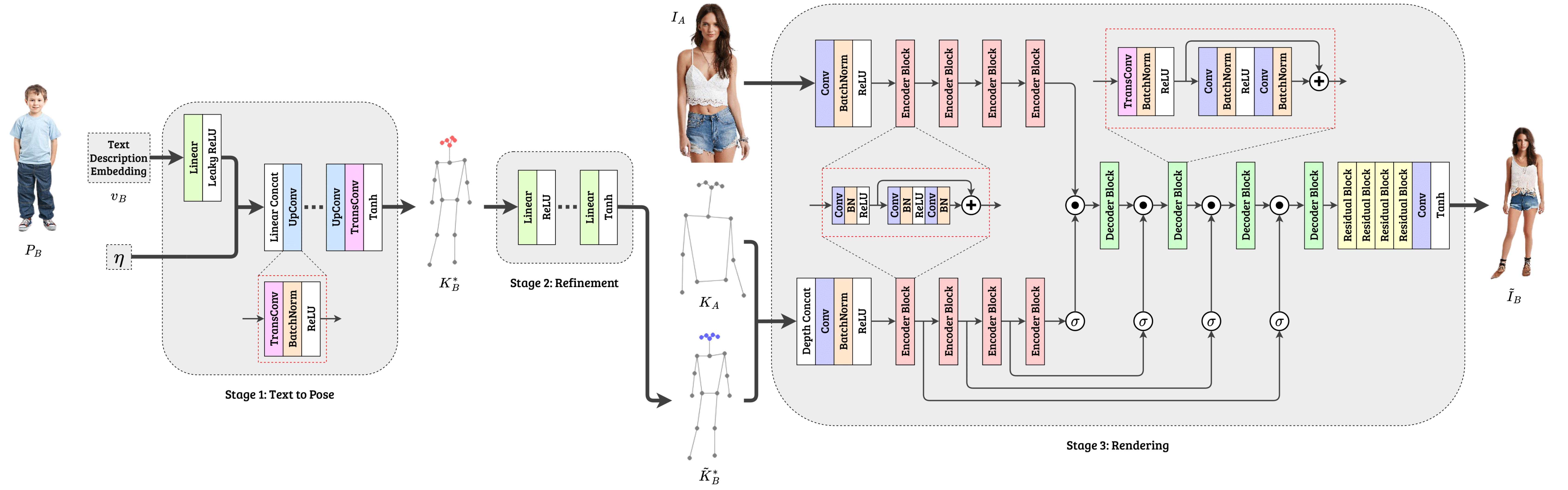
Table 2. Performance of pose transfer algorithms for real-world targets.

Pose Generation Algorithm	SSIM	IS	DS	PCKh	GCR	LPIPS (VGG)	LPIPS (SqzNet)
Partially Text Guided (Ours)	0.696	2.093	0.990	0.84	1.000	0.262	0.155
Fully Text Guided (Ours)	0.695	2.171	0.991	0.85	1.000	0.263	0.157
Zhou et al.	0.615	2.891	0.931	0.52	1.000	0.271	0.182
PATN	0.677	2.779	0.996	0.64	1.000	0.294	0.183
Real Data	1.000	2.431	0.984	1.00	1.000	0.000	0.000

Resources

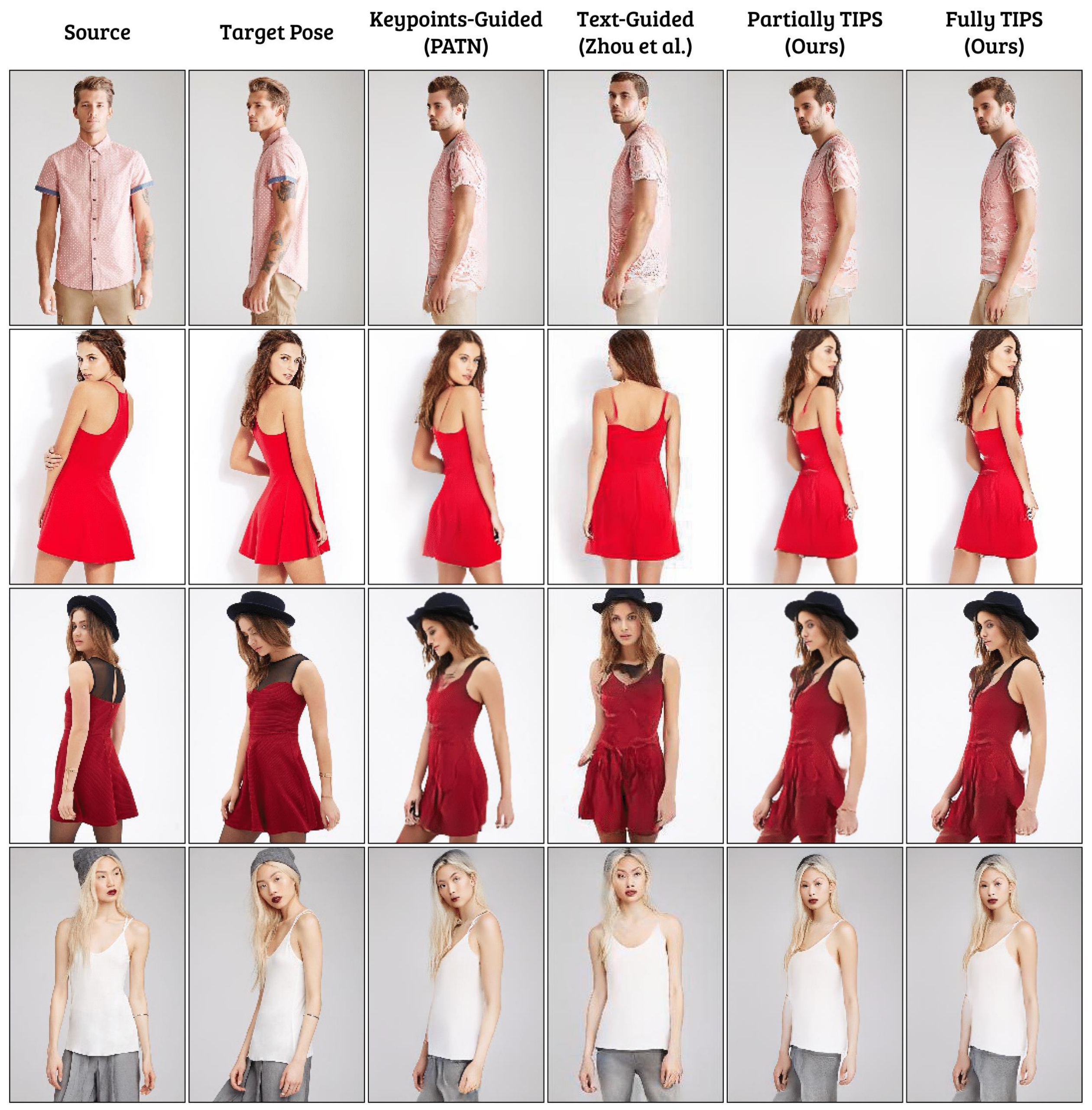
Visit the project website at https://prasunroy.github.io/tips for additional resources.

Pipeline Architecture

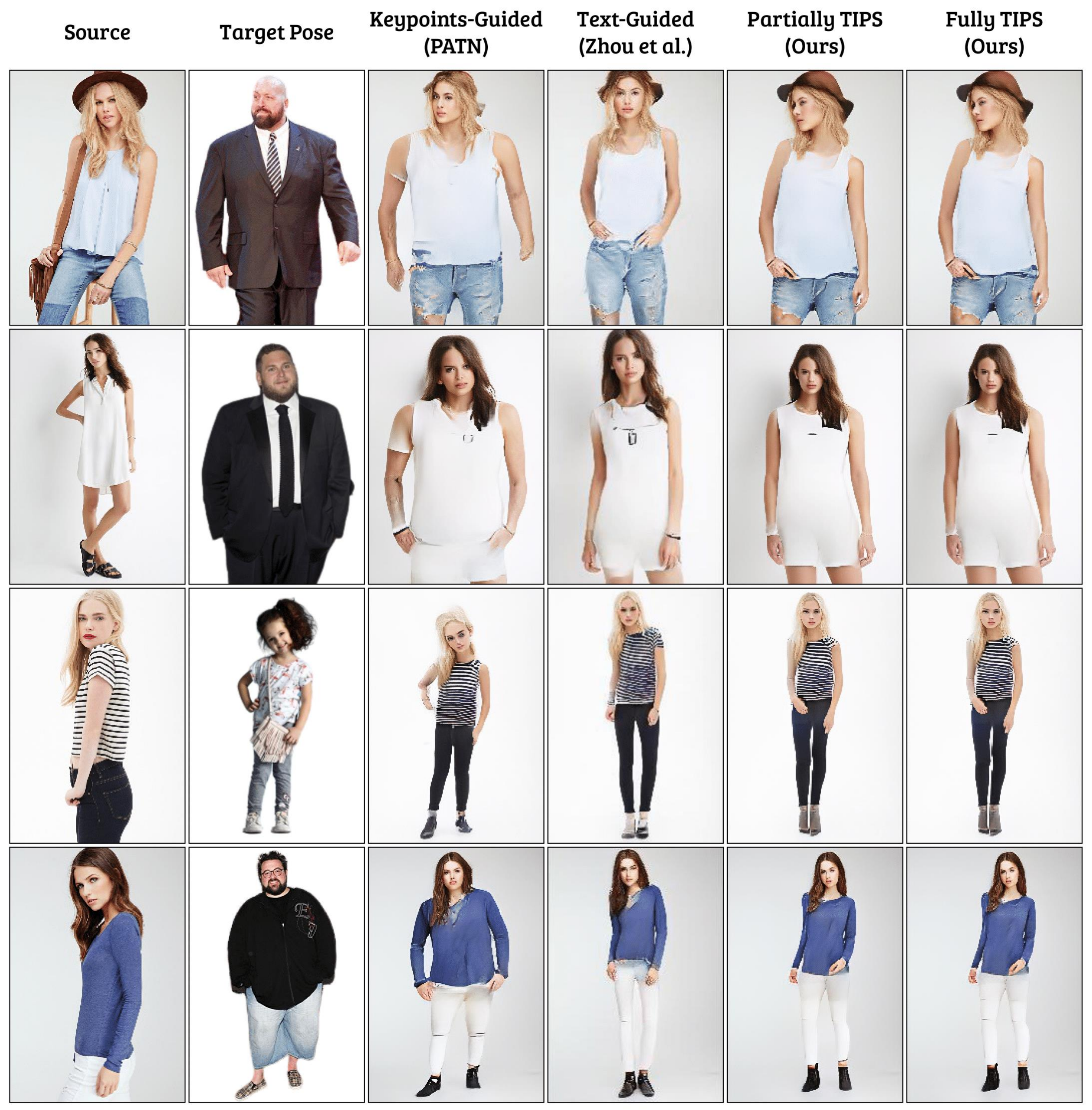


The pipeline is divided into three stages. In stage 1, we estimate the target pose keypoints from the corresponding text description embedding. In stage 2, we regressively refine the initial estimation of the facial keypoints and obtain the refined target pose keypoints. Finally, in stage 3, we render the target image by conditioning the pose transfer on the source image.

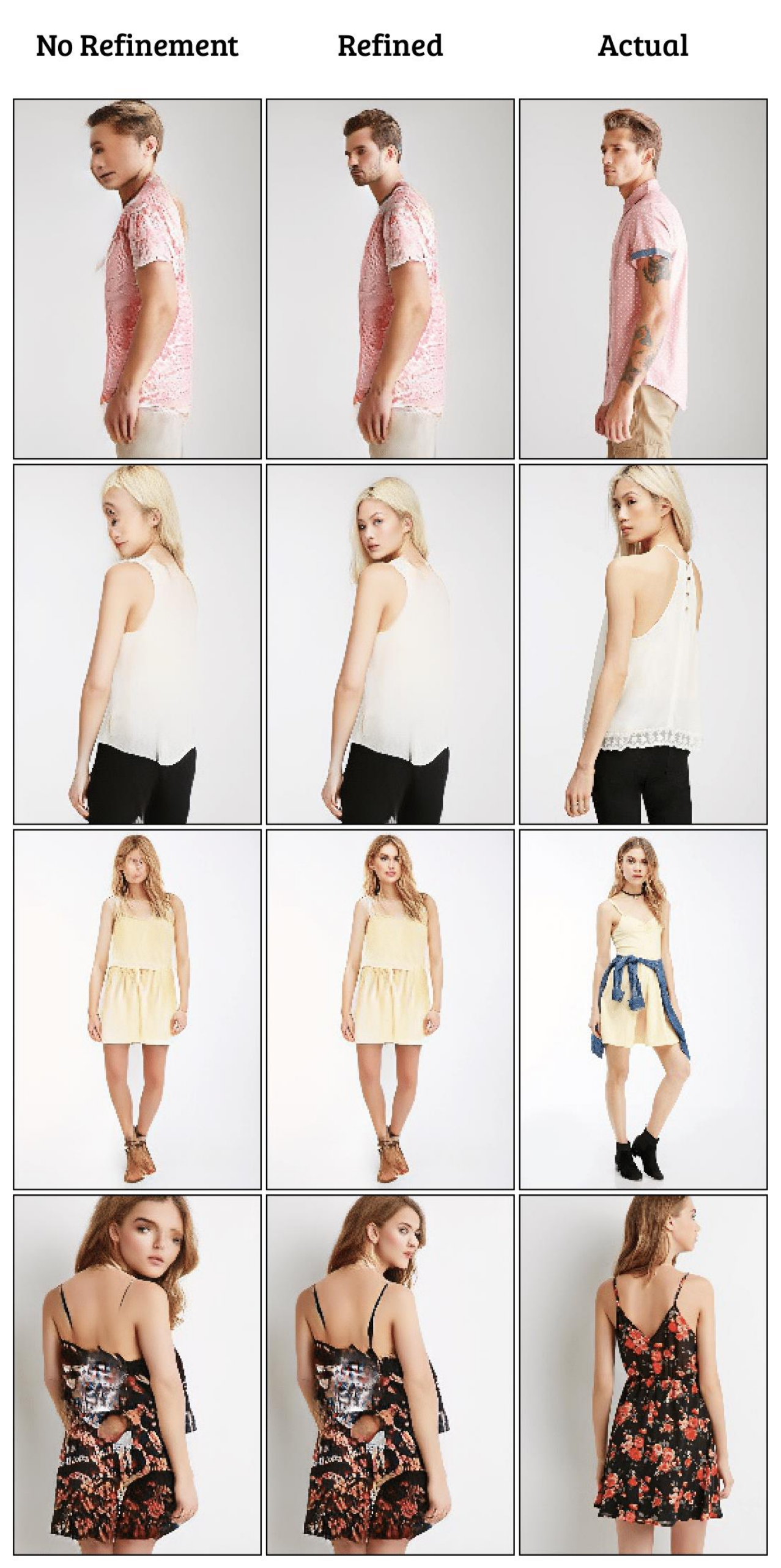
Visual Results



DeepFashion (Within distribution target pose samples)



Real World (Out of distribution target pose samples)



Effects of Face Refinement