Supplementary Material
for
TIPS: Text-Induced Pose Synthesis

Prasun Roy[1], Subhankar Ghosh[1], Saumik Bhattacharya[2],
Umapada Pal[3], and Michael Blumenstein[1]

[1] University of Technology Sydney, Australia
prasun.roy@student.uts.edu.au, subhankar.ghosh@student.uts.edu.au,
michael.blumenstein@uts.edu.au
[2] Indian Institute of Technology Kharagpur, India
saumik@ece.iitkgp.ac.in
[3] Indian Statistical Institute Kolkata, India
umapada@isical.ac.in
https://prasunroy.github.io/tips

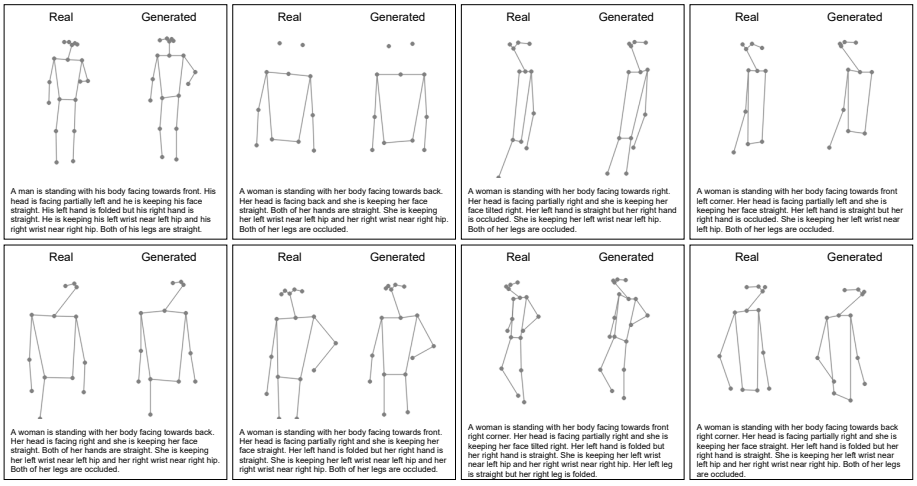

Fig. 1. Additional qualitative results of text to pose generation in stage 1. For each example, **Left:** Actual target pose (Ground Truth), **Right:** Generated pose conditioned purely on the textual description of the target pose, **Bottom:** Textual description of the target pose.

Fig. 2. Additional qualitative comparison among different pose transfer algorithms. Keypoint-guided methods tend to produce structurally inaccurate results when the physical appearance of the target pose reference significantly differs from the condition image. This observation is more frequent for the *out of distribution* target poses. The proposed text-guided technique successfully addresses this issue while retaining the ability to generate visually decent results close to the keypoint-guided baseline.

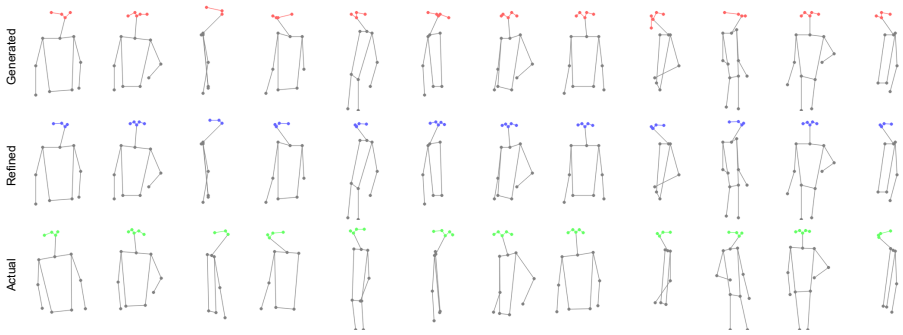

Fig. 3. Additional qualitative results of regressive refinement in stage 2. The refinement is performed specifically on the facial keypoints (marked with color). **Top:** Estimated keypoints from textual description in stage 1. **Middle:** Refined keypoints in stage 2. **Bottom:** Actual keypoints (Ground Truth).

Fig. 4. Additional qualitative results with and without refinement. **Top:** Images generated without refinement. **Middle:** Images generated with refinement. **Bottom**: Actual target images (Ground Truth). The regressive refinement (stage 2) significantly improves the final generation quality by correcting the spatial coordinates of the keypoints estimated from textual description (stage 1).

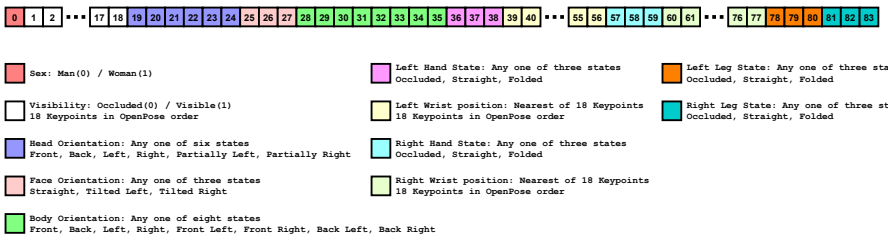

Fig. 5. The layout of the many-hot encoding vector in the proposed DF-PASS dataset.

Fig. 6. Text-assisted 180° interpolation in standing pose using the proposed method.