Single-Shot Freestyle Dance Reenactment Supplementary

A. Additional results

Body diversity. As mentioned, explicit augmentations encourage diverse body structure preservation. Fig. 1 showcases this aspect, where two individuals are chosen with distinctively different body structures. The semantic maps of both individuals are shown in the first row, while the generated semantic map for the same pose is shown in the second row. The individuals are overlaid in column (c) for clarity.

Sample results. Additional results are provided in Fig. 3 for both ”simple” and ”challenging” target images, over different poses. In all cases, realistic samples are rendered.

Interchangeable backgrounds. Generating a blending mask is an integral part of the method, as it enables embedding the generated person into any background. Fig. 4 demonstrates this ability. As seen in column (c), by embedding the rendered person back into the inpainted source video, the shadows of the original dancer complement the naturalness of the rendered person.

B. Additional Comparison

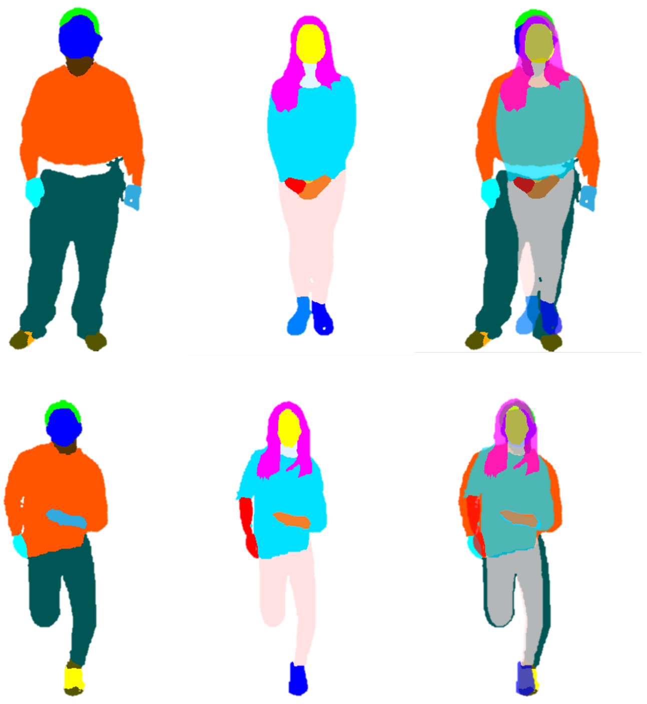
Comparison with Liquid-GAN [2] is presented in Fig. 2. Compared to [2], our biggest advantage is natural motion, which cannot be conveyed here. As shown in Fig. 2, our method also surpasses in terms of resolution, appearance, pose, and background replacement.

C. Additional implementation details

The P2B and B2F networks are trained with the *ADAM*[1] optimizer applying a learning rate of 0.0002 and $(\beta1, \beta2) = (0.5, 0.999)$. The P2B is trained for 280 epochs, with a batch size of 128, while the B2F is trained for 60 epochs, with a batch size of 32. The Face Refinement network is trained with the same optimizer, a learning rate of 0.0001, $(\beta1, \beta2) = (0.5, 0.999)$, for 40 epochs and a batch size of 256.

D. Limitations

Our method is driven by pose representations, and conditioned over a semantic map of the target person. As previous methods, ours as well suffers from a strong dependency on the quality of the detected driving pose, though is somewhat robust to the conditioned semantic map (hence capable of handling ”challenging” targets).



(a) (b) (c)

Figure 1. Body structure diversity example. For the same driving pose, two generated individuals are evaluated. The body structure, as captured by the semantic segmentation of the target images (row 1) for the first (a) and second (b) person, can be see to be distinct, as emphasized by overlaying one over the other (c). The distinction in body structure can be seen to be maintained in the corresponding rendered images (row 2).

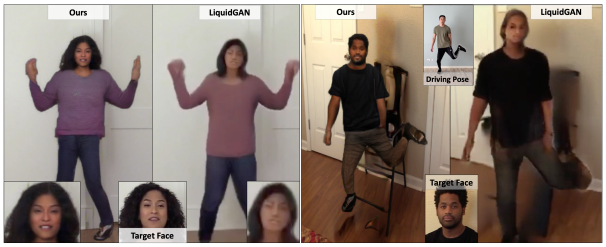


Figure 2. Ours vs. LiquidGAN. (L) Easy, (R) challenging targets.

Body structure preservation is an important aspect of

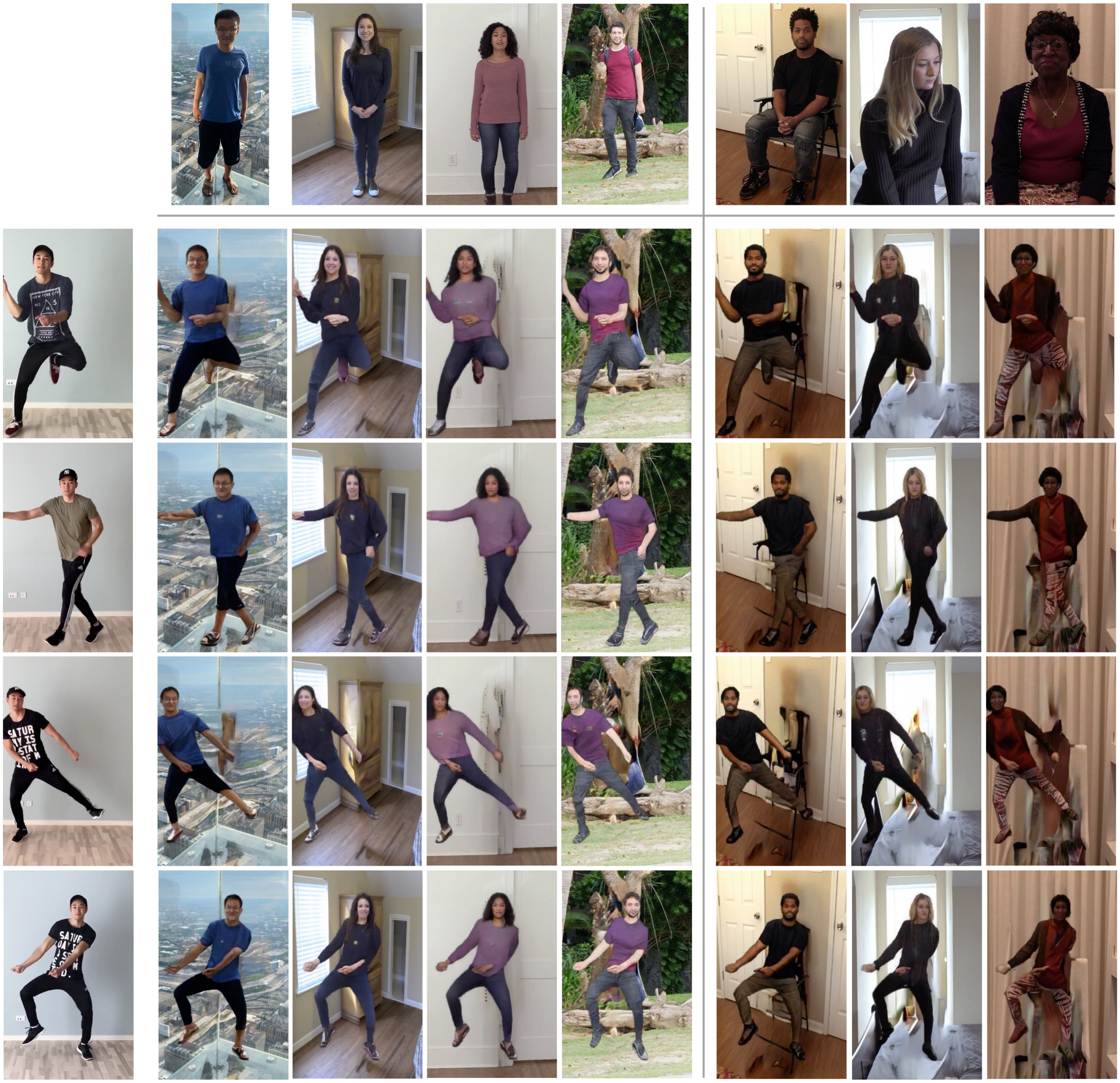


Figure 3. Sample results. Four ”simple” and three ”challenging” targets are shown. In all cases, realistic samples are rendered for a diverse set of appearances and poses. Additional results can be seen in the accompanying video. Note that the facial expression is transferred from the target image, rather than from the driving image.

dance reenactment, and receives significant attention in this work. Although this method is able to preserve some body structure, it is still constrained by the strong bias that accompanies datasets used to train the different networks, specifically the Pose2Body network.

The rendered blending mask enables to seamlessly blend the generated person into any given background, yet does not provide a complete solution for all environmental surroundings, such as shadows. A partial resolution for this gap is using the inpainted source video as the background, as seen in Fig. 4(c) and in the accompanying video.

E. P2B ablation experiment.

The ablation experiment for the P2B network is presented in Fig. 7. We highlight dominant discrepancies by a green square for our result and a red square for each ablation case.

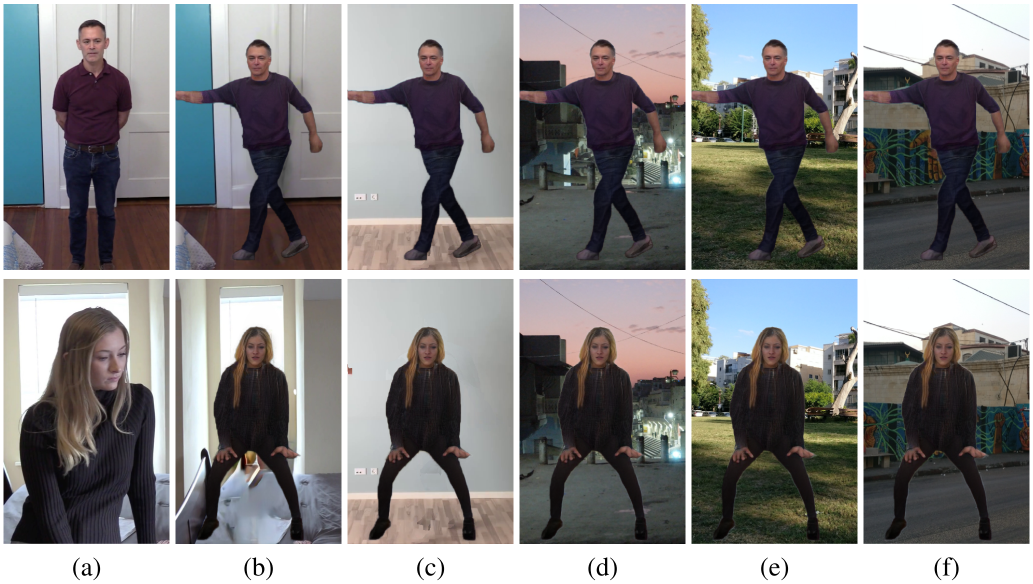


Figure 4. Interchangeable backgrounds. The generated blending mask is used to seamlessly embed the rendered person into any given background. (a) Target image, (b) embedded into the inpainted target background (c) embedded into the inpainted driving video background (residual shadows complement the naturalness of the embedded person), (d)-(f) embedded into various backgrounds.

F. Quantitative ablation

We focus on a qualitative ablation for the following reasons: (1) As the main objective is rendering a novel person, real dance generation does not have a ground-truth, making majority of the metrics irrelevant (e.g. disentangling the body structure from the driving pose is not relevant, resulting in deceptively better results for the ablation case), (2) numerical metrics often hide the real impact of losses trade-offs. As an example, we achieve better LPIPS if we do not use any face-related losses, as the addition of a face-related loss adds conflicting considerations. However, face appearance is very important in human perception. Nevertheless, quantitative results are presented in Tab. 1. As expected, it shows a trade-off between the losses, e.g., removing the face-related losses hurts face perception significantly, while slightly improving other metrics.

G. Inference time

Inference time considerations mainly focus on mitigating bottlenecks and maximum parallelization. The main bottlenecks are currently the DensePose and B2F networks' run-time. To achieve real-time inference, we would either remove DP, or employ DP on a low-resolution image. Reducing the B2F run-time could be achieved by a range of optimizations, such as reducing channel number, or converting ResSPADE blocks to lighter ResBlocks (e.g. MobileNetV3). This results with the sequence of (1) OP+DP, (2) P2B, (3) B2F, (4) FR (the rest is done once per person, and could be pre-processed). As we do not employ any temporal components, each of the 5 networks could run in parallel on 5 GPUs (after passing the first 4 frames). This would bring us to approx. (1) 41ms, (2) 20ms, (3) 20ms, (4) 30ms, where (1) is the limiting factor, resulting in 24FPS (can be improved by adding an additional GPU for OP), with a latency of 111ms.

H. Region refinement

The face refinement utilized a network trained specifically on faces to improve quality and appearance. In a similar manner to face refinement, it is possible to add losses emphasizing each part of interest (e.g. hands, shirt, pants), utilizing a specific network (e.g. trained on hands) or a general one (e.g. ImageNet). This is already done implicitly through the pre-trained encoder, yet explicit losses (as done for the face part) can provide additional improvement.

References

[1] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2016.

[2] Wen Liu, Zhixin Piao, Min Jie, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

Figure 5. ”Simple” targets used for human preference survey and visual comparison.

Method	SSBS ↑	SSIS ↑	DPBS ↑	DPIS ↑	LPIPS ↓ (VGG)	LPIPS ↓ (SqzNet)	SSIM ↓	FID ↓
(P2B) No Squeeze/Stretch	0.898	**0.220**	0.926	**0.514**	-	-	-	-
(P2B) No Accurate DP	**0.902**	0.218	0.927	0.500	-	-	-	-
(P2B) No DP	0.869	0.197	0.884	0.460	-	-	-	-
(P2B) No Fingers/DP	0.869	0.197	0.884	0.460	-	-	-	-
(B2F) No FR	0.873	0.208	0.896	0.468	0.378	0.299	0.133	**70.880**
(B2F) No Mask	0.873	0.216	0.891	0.458	0.379	0.300	0.135	74.503
(B2F) No Fingers	0.863	0.208	0.897	0.467	0.375	0.296	0.130	73.715
(B2F) No Face-loss/LR	0.873	0.217	0.896	0.465	**0.373**	0.293	0.128	77.032
Ours	**0.902**	0.218	**0.928**	0.500	0.375	**0.283**	**0.116**	83.95

Table 1. Quantitative ablation.

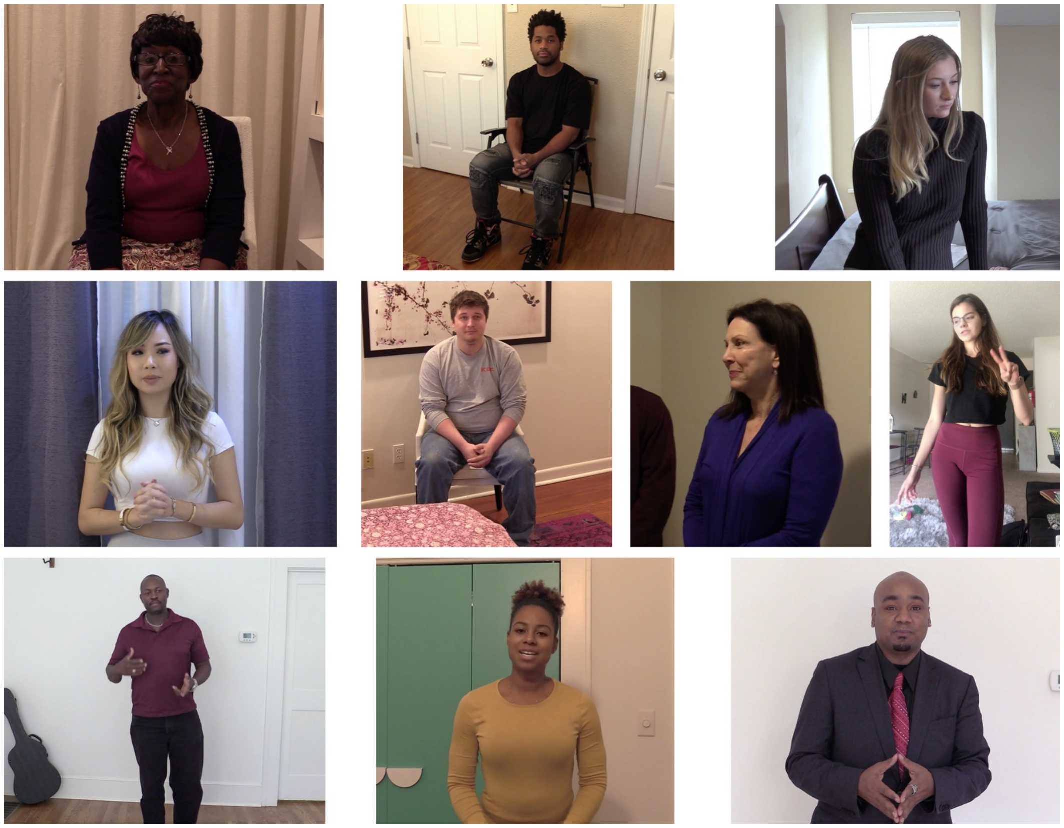


Figure 6. ”Challenging” targets used for visual comparison.

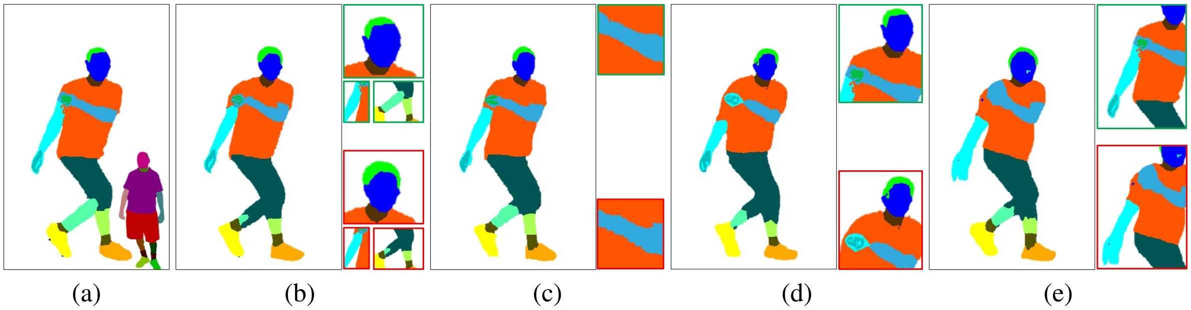


Figure 7. P2B ablations. (a) Our result and the target parsing (scaled down). The following are various variants. In red, a zoom in version, and in green the same zoom applied to the output of the full method. (b) No squeezing and stretching of the input/output parsing (*body structure, hair, and clothing less consistent*), (c) a less accurate version of DensePose is used (*boundary artifacts*), (d) DensePose is not used as input (*increased limbs artifacts, instability in body structure*), (e) no DP and no hand/finger labels (*enormous arms*).