Cross-view and Cross-pose Completion for 3D Human Understanding
Supplementary Material

Matthieu Armando Salma Galaaoui Fabien Baradel Thomas Lucas
Vincent Leroy Romain Brégier Philippe Weinzaepfel Grégory Rogez
NAVER LABS Europe
https://europe.naverlabs.com/ComputerVision/CroCoMan

This supplementary material presents additional results on the human texture estimation task (Section 1), training cost information (Section 2) and qualitative results of the pre-training objective as well as of several downstream tasks (Section 3). In addition, we also include a video showing reconstructions using our pre-training method.

1. Human texture estimation

Our pre-training objective has some similarities with the task of novel-view synthesis. Given an observation of a person (the reference image), and some information about a target pose and viewpoint (the masked target image), the network is trained to reconstruct an image of the person from said viewpoint. In order to evaluate this particular facet of human understanding, we compare different pre-training strategies on the task of human novel-view generation. More particularly, we tackle human texture generation from a single image, following the experimental setup of TexFomer [5]. They define a key, query and value images which are partly pre-computed, and partly based on the input image. These images are encoded at different scales using 3 CNNs, then transformer layers perform multi-headed attention at different scales. Resulting features are merged through another CNN. We modify their code, replacing their whole network with our ViT-based encoder-decoder architecture. The value image is discarded, and encoder weights are fine-tuned independently for key and query images. The network is trained to return a single RGB texture. This adaptation is a bit naive, but our goal is mainly to compare different pre-training methods on a different task, that leverages both encoder and decoder of the pre-trained network. We follow the TexFormer experimental setup in terms of hyperparameters, datasets, and metrics. Results for different network initializations are reported in Figure 1. For MAE, we randomly initialize the decoder weights. Pre-training the model does help a lot: both CroCo and MAE provide a significant boost. CroCo performs slightly better, which is probably due in part to the pre-trained decoder. CroCo-Body outperforms both CroCo and MAE.

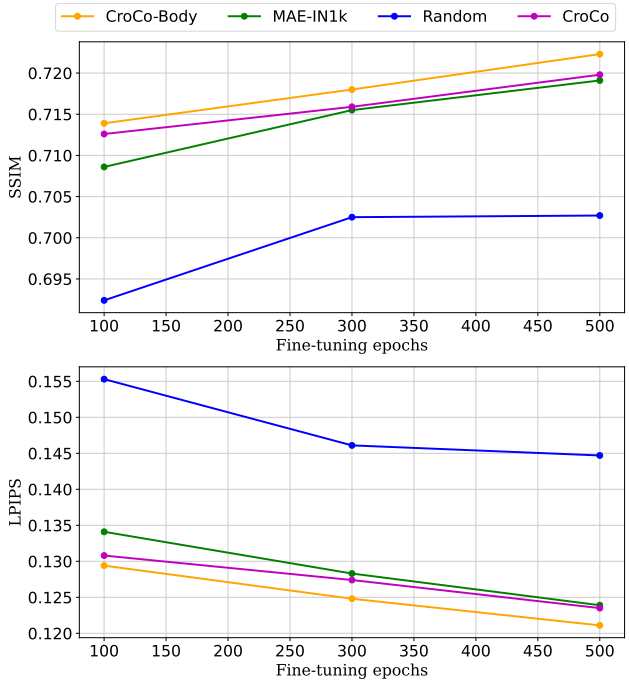


Figure 1. **Evaluation scores of various pre-trained models on the texture estimation task of TexFormer [5]**, at different fine-tuning stages. From left to right, we report SSIM↑ (structural similarity index) and LPIPS↓ [6] metrics. All models return a single RGB texture.

2. Training time

In this section, we give timings necessary for pre-training and fine-tuning our models. Pre-training the CroCo-Body model takes about 8 days on 4 NVIDIA A100 GPUs. Fine-tuning it on a single A100 takes about half a day per downstream task. As for the CroCo-Hand model, pre-training on 4 V100 GPUs requires 2.25 days, and fine-tuning on a single V100 GPU takes about 8 hours per downstream task.

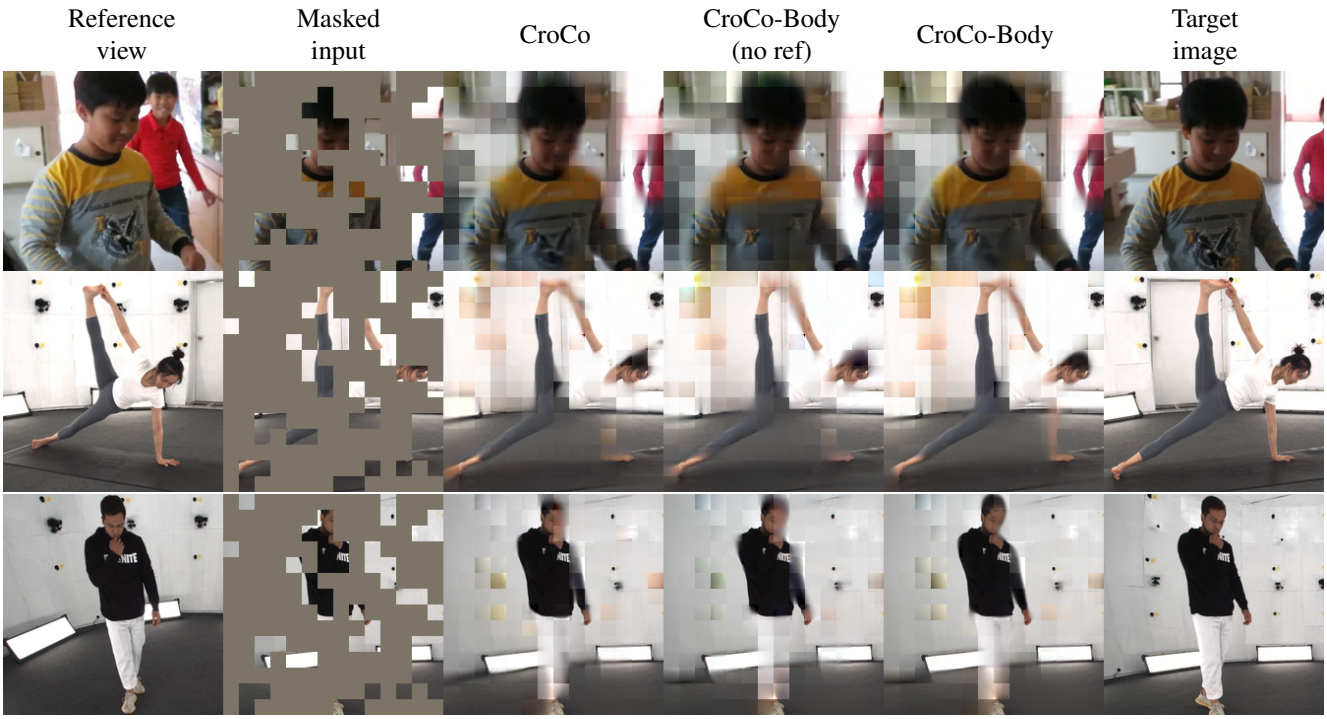


Figure 2. **Completion examples on cross-view (*i.e.* multi-view) pairs** from the Mannequin Challenge dataset [2] (first row) and the GeneBody dataset [1] (last two rows). CroCo-Body (no ref) stands for our model evaluated on the masked input, and a reference view set to zero (*i.e.*, a fully-black image).

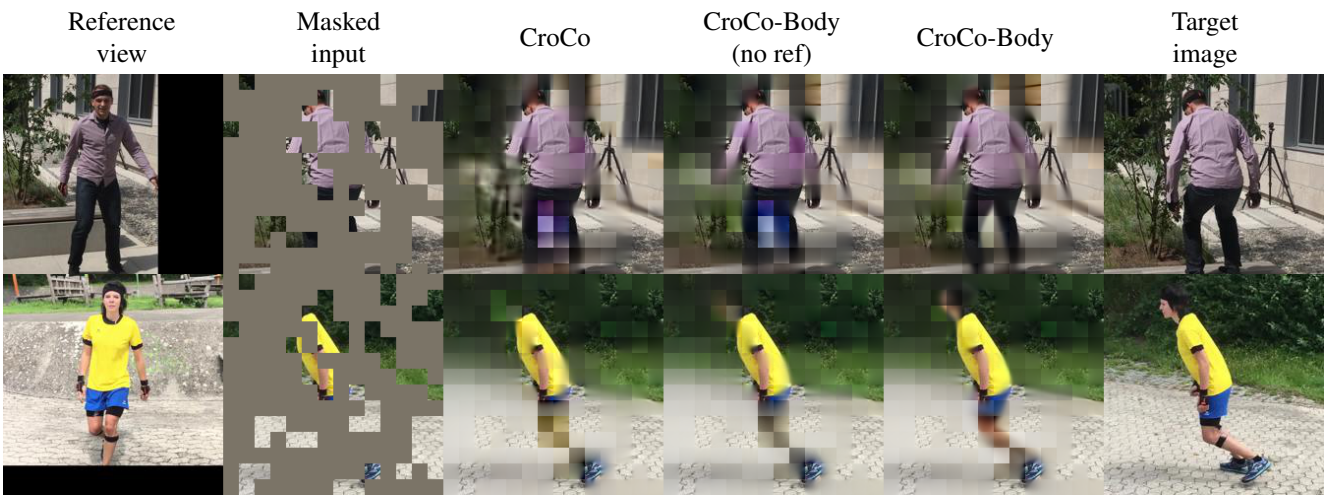


Figure 3. **Completion examples on cross-pose (*i.e.* temporal) pairs** from 3DPW [4] validation set (unseen during pre-training). CroCo-Body (no ref) stands for our model evaluated on the masked input, and a reference view set to zero (*i.e.*, a fully-black image).

3. Qualitative results

3.1. Pre-training

CroCo-Body. We illustrate the pre-training task of CroCo-Body on both cross-view and cross-pose pairs in Figures 2 and 3 respectively, with data never seen by the model during pre-training. We report predictions of CroCo-Body using either the reference image or a reference image entirely black ('no ref'), to ablate the cross-image completion capabilities of the decoder. CroCo tends to recover detailed patterns on relatively flat surfaces, such as the t-shirt logo on the first row of Figure 2. It lacks prior knowledge about humans however, and struggles to reconstruct the left arm on the second row. In contrast, CroCo-Body produces a

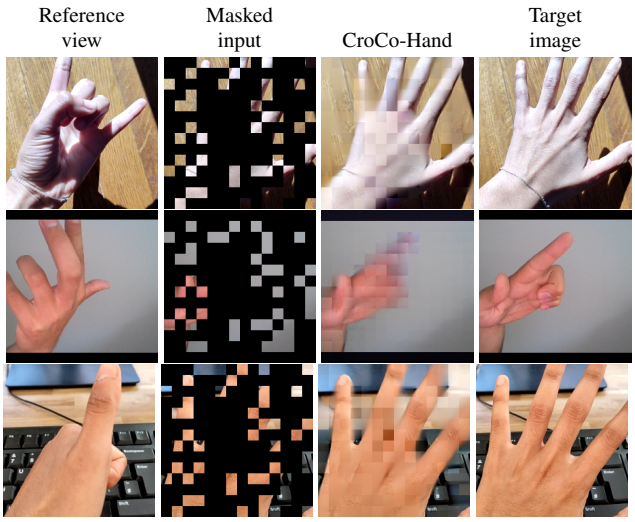


Figure 4. **Completion examples of CroCo-Hand on unseen cross-pose (*i.e.* temporal) pairs**, in indoor scenes.

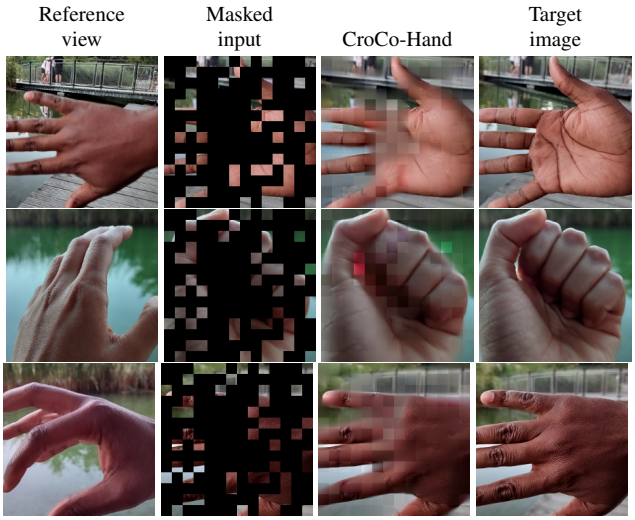


Figure 5. **Completion examples of CroCo-Hand on unseen cross-pose (*i.e.* temporal) pairs**, in outdoor scenes.

sharper arm reconstruction, which may be attributed to its human-specific pre-training and the ability to leverage the reference view. A similar effect is visible on the reconstruction of the head in the last row.

For cross-pose pairs (Figure 3), we observe that completions of CroCo are similar to the ones of CroCo-Body with no reference image. This suggests that CroCo benefits little from cross-image attention, being specifically trained to exploit static stereoscopic pairs only. CroCo-Body on the other hand seems able to recover information from the reference image about the lower-body garments even though they are heavily occluded in the masked target in both examples, and achieves a better completion of the masked image.

CroCo-Hand. We illustrate the pre-training task of CroCo-Hand on unseen cross-pose pairs in indoor and outdoor scenes in Figures 4 and 5, respectively. We tested CroCo-Hand on internal images which have never been seen during the pre-training stage. We observe that CroCo-Hand learned the structure of a human hand such as shown in Figure 4 where it reconstructs a pointed index finger from a small handful of visible palm patches. CroCo-Hand also performs well on outdoor images such as shown in Figure 5, despite the fact the pre-training is done integrally using data captured in labs. It is also interesting to notice that CroCo-Hand also generalizes well to different skin tones.

Keypoints supervision. We give here more detailed information about the keypoints supervision used for the pre-training ablation in Section 4.2 and Table 3 of the main paper. We select the set of 13 keypoints used in PennAction [7]. For each pre-training image, we generate a 13-channels keypoint heatmap where each keypoint is represented as a Gaussian with $\sigma = 8$ pixels. Figure 6 illustrates the task on a simple example. During pre-training, the encoded image is passed through a simple prediction head that is trained to predict the heatmaps with a simple binary cross-entropy loss. Ground-truth keypoints are weighted according to a confidence parameter (0 for missing keypoints). When pre-training with both objectives (Table 3 of the main paper, last row), we train the keypoints prediction on the encoded reference image, that is fully visible.

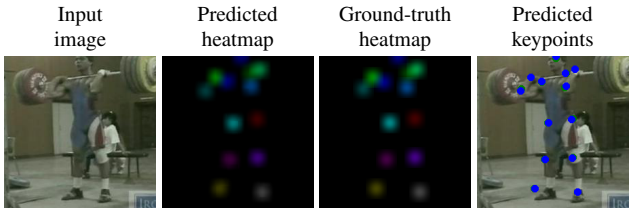


Figure 6. **Visualization of the supervised pretext task used for the ablation in Table 3 of the main paper**. The right shows the position of predicted keypoints (blue) obtained with a simple argmax on the predicted heatmap, on top of ground truth keypoints (green). Heatmaps have been artificially converted to 3-channels images for visualization purpose.

3.2. Downstream results

We now show some visualizations of the different downstream tasks that we evaluate on. Figures 7 and 8 show results on regression tasks (DensePose and body/hand mesh recovery, respectively), while Figure 9 shows results on the grasp classification task.

References

[1] Wei Cheng, Su Xu, Jingtan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. Generalizable neural performer: Learning robust radiance

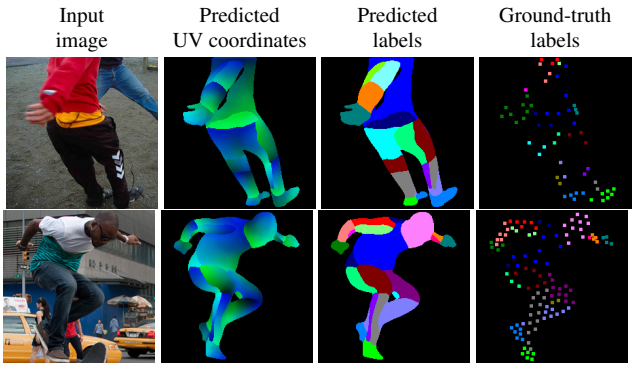
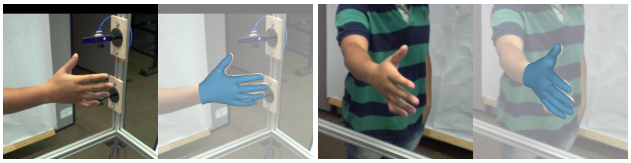


Figure 7. **Qualitative results of CroCo-Body on the DensePose task** on the COCO dataset. The sparse ground-truth labels used for training and evaluation are dilated here for visualization purposes.

(a) **Results on the body mesh recovery task** on 3DPW [4].



(b) **Results on the hand mesh recovery task** on HanCo [8].

Figure 8. **Qualitative examples of our models on the two mesh recovery tasks.** Each pair shows the input image, and the output of CroCo-Body (top) or CroCo-Hand (bottom), overlaid on the image.

Figure 9. **Qualitative examples of our models on the grasp classification task** on GUN-71 [3]. For the images on the top row, we show below the ground-truth class as well as the top 3 prediction made by our model.

fields for human novel view synthesis. *arXiv preprint arXiv:2204.11798*, 2022. 2

[2] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019. 2

[3] Grégory Rogez, James S Supancic, and Deva Ramanan. Understanding everyday hands in action from rgb-d images. In *ICCV*, 2015. 4

[4] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2, 4

[5] Xiangyu Xu and Chen Change Loy. 3D human texture estimation from a single image with transformers. In *ICCV*, 2021. 1

[6] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1

[7] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 3

[8] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In *GCPR*, 2021. 4