Leveraging Photometric Consistency over Time for Sparsely Supervised Hand-Object Reconstruction

Yana Hasson[1,2] Bugra Tekin[4] Federica Bogo[4]
Ivan Laptev[1,2] Marc Pollefeys[3,4] Cordelia Schmid[1,5]

[1]Inria, [2]Département d'informatique de l'ENS, CNRS, PSL Research University
[3]ETH Zürich, [4]Microsoft, [5]Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK

Our main paper described a method for joint reconstruction of hands and objects, and proposed to leverage photometric consistency as an additional source of supervision in scenarios where ground truth is scarce. We provide additional details on the implementation in Section A, and describe the used training and test splits on the HO-3D dataset [2] in Section B. In Section C, we detail the cyclic consistency check that allows us to compute the valid mask for the photometric consistency loss. Section D provides additional insights on the effect of using the skeleton adaptation layer.

A. Implementation details

Architecture. We extract image features from the last layer of ResNet18 [3] before softmax. We regress in separate branches 6 parameters for the global object translation and rotation, 3 parameters for the global hand translation, and 28 MANO parameters which account for global hand rotation, articulated pose and shape deformation. The details of each branch are presented in Table 1.

Training. All models are trained using the PyTorch [8] framework. We use the Adam [6] optimizer with a learning rate of $5 \cdot 10^{-5}$. We initialize the weights of our network using the weights of a ResNet [3] trained on ImageNet [10]. We empirically observed improved stability during training when freezing the weights of the batch normalization [5] layer to the weights initialized on ImageNet.

Branch	Input shape	Output shape	ReLU
`Object pose regressor`	512	256	✓
	256	6	
`Hand translation regressor`	512	256	✓
	256	3	
`Hand pose and shape regressor`	512	512	✓
	512	512	✓
	512	28	

Table 1. **Architecture of the Hand and Object parameter regression branches.** We use fully connected linear layers to regress pose and shape parameters from the $512-$dimensional features.

We pretrain the models on fractions of the data without the consistency loss. As an epoch contains fewer iterations when using a subset of the dataset, we observe that a larger number of epochs is needed to reach convergence for smaller fractions of training data. We later fine-tune our network with the consistency loss using a fixed number of 200 epochs.

Runtime. The forward pass runs in real time, at 34 frames per second on a Titan X GPU.

B. HO-3D subset

In Sec. 4.3, we work with the subset of the dataset which was first released. Out of the 68 sequences which have been released as the final version of the dataset, 15 have been made available as part of an earlier release. Out of these, we select the 14 sequences that depict manipulation of two following objects: the mustard bottle and the cracker box. The train sequences in this subset are the ones named SM2, SM3, SM4, SM5, MC4, MC6, SS1, SS2, SS3, SM2, MC1, MC5. When experimenting with the photometric consistency, we use SM1 and MC2 as the two test sequences. When comparing to the baseline of [2], we use MC2 as the unique test sequence.

C. Cycle consistent visibility check

Our consistency check is similar to [7, 4].

Following the notation of Sec. 3.1, let us denote the flow warping the estimated frame $I_{t_{ref}+k}$ into the reference one $I_{t_{ref}}$ by $W_{t_{ref}+k \rightarrow t_{ref}}$. Similarly, we compute a warping flow in the opposite direction, from the refer-

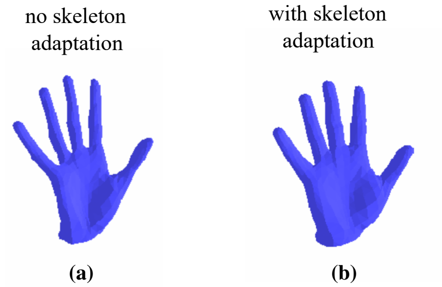

Figure 1. Predicted shape deformations in the **(a)** absence and **(b)** presence of the skeleton adaptation layer on the FPHAB dataset.

ence frame to the estimated one: $W_{t_{ref} \to t_{ref}+k}$. Given the mask $M_{t_{ref}}$ obtained by projecting $V_{t_{ref}}$ on image space, we consider each pixel $p \in M_{t_{ref}+k}$. We warp p into the reference frame, and then back into the estimated one: $\tilde{p} = W_{t_{ref}+k \to t_{ref}}(W_{t_{ref} \to t_{ref}+k}(p))$. If the distance between p and $\tilde{p}$ is greater than 2 pixels, we do not apply our loss at this location. On FHB, when using 1% of the data as reference frames, this check discards 3.3% of $M_{t_{ref}+k}$ pixels.

D. Skeleton Adaptation

The defined locations for the joints do not exactly match each other for the FPHAB [1] dataset and the MANO [9] hand model. As shown in Table 2 of our main paper, we observe marginal improvements in the average joint predictions using our skeleton adaptation layer. This demonstrates that MANO [9] has already the ability to deform sufficiently to account for various skeleton conventions. However, these deformations come at the expense of the realism of the reconstructed meshes, which undergo unnatural deformations in order to account for the displacements of the joints. To demonstrate this effect, we train a model on the FPHAB [1] dataset, without the linear skeleton adaptation layer, and qualitatively compare the predicted hand meshes with and without skeleton adaptation. We observe in Fig. 1(a) that, without skeleton adaptation, the fingers get unnaturally elongated to account for different definitions of the joint locations in FPHAB and MANO. As shown in Fig. 1(b), we are able to achieve higher realism for the reconstructed meshes using our skeleton adaptation layer.

References

[1] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[2] Shreyas Hampali, Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Ho-3d: A multi-user, multi-object dataset for joint 3d hand-object pose estimation, 2019. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[4] Junhwa Hur and Stefan Roth. MirrorFlow: Exploiting symmetries in joint optical flow and occlusion estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015. 1

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014. 1

[7] Natalia Neverova, James Thewlis, Riza Alp Guler, Iasonas Kokkinos, and Andrea Vedaldi. Slim densepose: Thrifty learning from sparse annotations and motion cues. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[8] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *Advances in Neural Information Processing Systems Autodiff Workshop*, 2017. 1

[9] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2017. 2

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015. 1