1 Implementation Details

SeqHAND-Net is structured as ResNet-50 [2] with the latent dimension of 1536, which is extended with a Conv-LSTM layer with 256-dimensional hidden states and a fully-connected (FC) layer to induce 26 variables including pose θ, shape β and view parameters r, t, s in our proposed structure. The SeqHAND-Net replaces the average-pooling layer and the FC layer of original ResNet50 with a Conv-LSTM layer and a FC layer. This may put our model in an advantageous position over the baseline model with more computational resource, but when the baseline is trained with ResNet101, its performance has not been improved as much as our extension has allowed, as shown in Table 2 in the main paper. The dimension of the latent and the hidden variables of the proposed structure are empirically chosen for its competitiveness against the baseline model though of the extension for temporal feature exploitation. SeqHAND-Net is resultantly composed of 43.2M trainable parameters. A pytorch impelementation [1] is used for MANO hand model. For fine-tuning for synthetic-real domain transition, we have set only the Conv-LSTM layer detached from further learning, letting the last output FC layer be tuned along with the encoder.

Training Loss for SeqHAND Dataset. The criterion for pre-training for sequential synthetic hand motion image dataset is a combination of re-projected 2D joint loss L_{2D}, a 3D joint loss L_{3D}, a temporal consistency loss L_{temp}, a loss for camera parameters L_{cam} and the mask fitting loss L_{mask}:

$$L = \lambda_{2D}L_{2D} + \lambda_{3D}L_{3D} + \lambda_{temp}L_{temp} + \lambda_{cam}L_{cam} + \lambda_{mask}L_{mask}. \quad (1)$$

Training Loss for Domain Transition to Real. We have utilized datasets of Stereo Benchmark [3] and FreiHand [4] for domain transfer of our trained network into real domain. The two datasets are differently annotated; STB datasets are only annotated with 2D and 3D joint locations while FreiHand dataset provides hand masks along with the 2D and 3D ground-truths. The loss of adaptation to real hand images is thus a combination of re-projected a 2D joint loss L_{2D}, a 3D joint loss L_{3D}, a hand mask loss L_{mask}, and a temporal loss L_{temp} :

$$L = \lambda_{2D}L_{2D} + \lambda_{3D}L_{3D} + \lambda_{temp}L_{temp} + \lambda_{mask}L_{mask}. \quad (2)$$

We have set the weights as $\lambda_{2D} = 5, \lambda_{3D} = 100, \lambda_{temp} = 100$ with $\lambda^{\theta}_{temp} = 2e^{-4}, \lambda_{cam} = 1$ and $\lambda_{mask} = 10$ for both pre-training with SeqHAND dataset and domain adaptation to real hand images.

References

1. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: CVPR (2019)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
3. Zhang, J., Jiao, J., Chen, M., Qu, L., Xu, X., Yang, Q.: A hand pose tracking benchmark from stereo matching. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 982–986. IEEE (2017)
4. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 813–822 (2019)