Generative Model-Based Loss to the Rescue: A Method to Overcome Annotation Errors for Depth-Based Hand Pose Estimation —SUPPLEMENTARY MATERIAL—

Jiayi Wang Franziska Mueller Florian Bernard Christian Theobalt
Max Planck Institute for Informatics, Saarbrücken, Germany

I. Projection of Model Gaussians

In this section we explain how our decoder layer "renders" the 3D model Gaussians into 2.5D representations. The h-th 3D model isotropic Gaussian $g_{\mu_h,\sigma_h}(x)$ is parameterized by the mean μ_h and the standard deviation σ_h. The camera-facing surface of the 3D Gaussian in the image plane is approximated by a depth value z_p that is associated with the projected 2D Gaussian $g_{\mu_p,\sigma_p}(x) = \Pi_K(g_{\mu_h,\sigma_h}(x))$, where

$$\mu_p = \frac{K \cdot \mu_h}{[\mu_h]_z}\,, \tag{1}$$

$$\sigma_p = \frac{\sigma_h f}{[\mu_h]_z}\,, \text{ and} \tag{2}$$

$$z_p = [\mu_h]_z - \sigma_p\,. \tag{3}$$

Here, K is the intrinsic camera matrix, f is the focal length of K, and $[\mu_h]_z$ is the z component of μ_h.

II. Balancing the Energy Terms

To find the weights λ, we used our implementation of the losses as a traditional model-based tracker on single images. We tune the weights so that the different terms evaluates to similar values in this setting and observe that these weights work for training as well. The following weights were used for all experiments: $\lambda_{\text{dissim}} = \lambda_{\text{collision}} = 0.6$, $\lambda_{\text{bone}} = 10^{-4}$, $\lambda_{\text{lim}} = 0.5$, $\lambda_{\text{joint}} = 8 \cdot 10^{-6}$.

III. Relevance of Pre-Processing

In the literature it is common practice to "crop around the hand" or to "crop the hand area using the ground truth joint locations" as a pre-processing step (e.g. [1], [4], [5]). However, such statements are oftentimes ambiguous as it do not precisely specified how the box is centered. We show here that the choice of the centering algorithm could hide the effects of annotation errors in cross-benchmark evaluations. Some previous works (e.g. [2], [3]) center their 3D bounding box at the average position of the *ground truth annotations* of the 3D joints, and then predict locations relative to this center. In a real application setting, ground truth joint positions are not available. Instead, one needs to *estimate* the hand location, which in turn may lead to inaccurate 3D bounding boxes and thereby results in offsets in the cropped depth images. As such, we argue that it is of crucial importance that a method can account for inaccurate localization, so that the predictions still yield the correct global 3D pose. However, we found that methods trained on 3D ground truth crops do not generalize to different depth offsets.

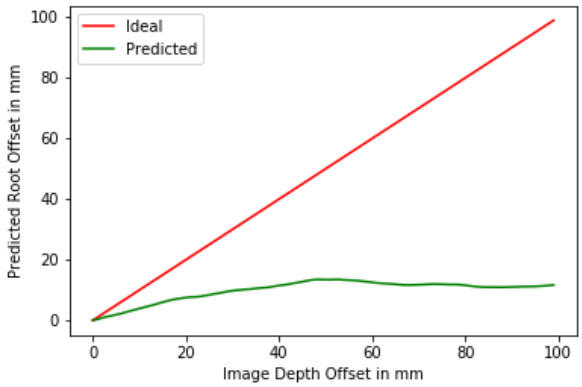

Fig. 1: **Depth Shift Test:** We plotted the shift in hand root position vs the mean offset in foreground depth values for a method trained on "Ground Truth" centered crop (**Predicted**). This show that such methods ignore global offsets in hand position relative to the crop center.

Fig. 1 shows that when we add a constant depth offset to the cropped image, the predicted hand root position does not change accordingly. This means that the predictor only learns to infer articulation *relative to the root rather than relative to crop center*. Thus, when running a cross-benchmark evaluation with such a pre-processing step for datasets with and without annotation errors, the errors (which are global offsets) are hidden since the crop

annotations are used to center the crop.

To mitigate this effect, we put our 3D bounding box center at the mean depth value in the image. Note that we hence only require a rough 2D crop at test time since the average depth value can be directly computed from the pixels. This information can be obtained more reliably than the average depth of the 3D ground truth joint positions. When evaluating across benchmarks with this pre-processing, the bias becomes apparent in that the predictor trained on HIM (which has annotation bias) consistently predicts the hand to be significantly behind the point cloud (see Fig. 2). We emphasize that our generative model-based loss, in combination with a slack radius, is able to correct this bias.

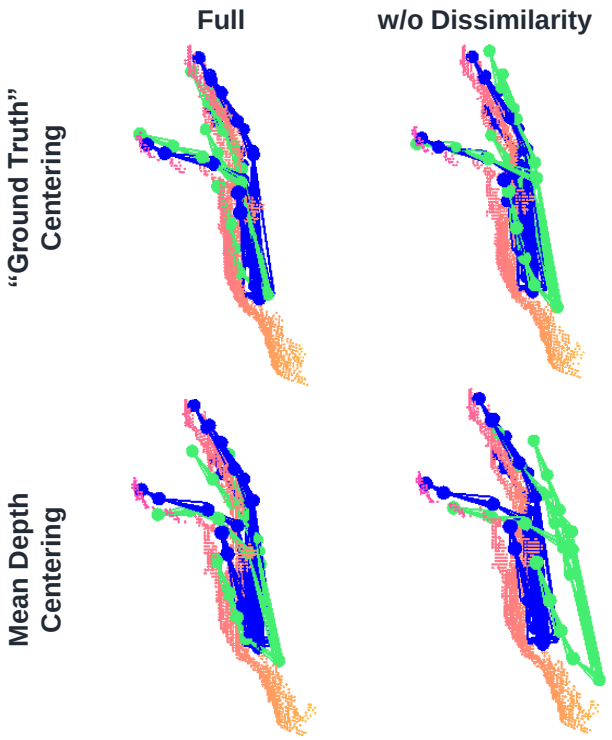

Fig. 2: **Hidden Bias: HIM to NYU.**. Using "Ground Truth" centering (Top) hides the annotation bias in the HIM dataset while our pre-processing step (Bottom) reveals the bias when dissimilarity loss is removed.

IV. Our HandID Dataset

Our HandID Dataset contains a total of 3,601 frames from 7 subjects that were acquired with the Intel SR300 sensor. The subjects were instructed to perform simple abduction, adduction, and flexion gestures while the camera recorded in a third-person view. The annotators were told to select 6 pixels for each depth image that correspond to the fingertip and wrist keypoints. In case the keypoints are occluded, the annotators were asked to estimate plausible 2D locations based on previous and future frames in the image sequence and to mark them as occluded. With that, for visible keypoints we obtain 3D annotations using the depth values of the pixels, and for occluded keypoints we obtain 2D image space annotations. See Fig. 3 for examples.

V. Visualization of the HIM biases

We present more visual results on the HIM dataset test set of [3] in Fig.4. Our method is trained only with the biased annotations provided in the HIM dataset. Each cell shows the camera view (**left**) and a novel view (**middle**) of the same prediction. Each cell additionally shows the prediction from a State-of-the-Art method (SotA) [3] visualized in the novel view (**right**). Note that our predictions (**green**) generally align much better to the depth images in the novel view than the annotations (**blue**). The SotA (**black**) methods also do not align well to the depth image as they learn to reproduce the bias by minimizing only the per-joint-errors during training.

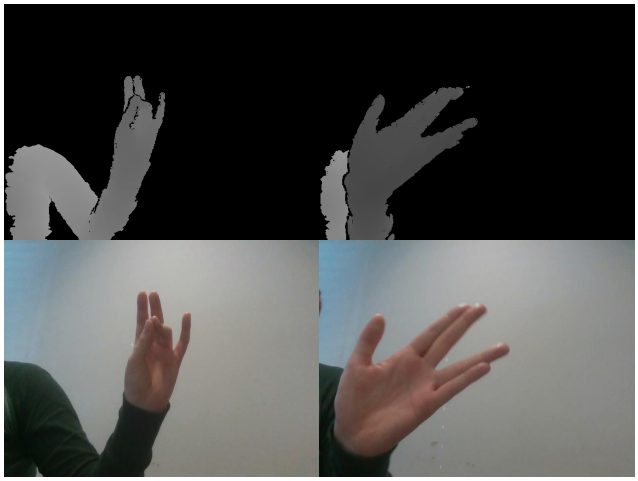

Fig. 3: **HandID Dataset:** We show two examples of the depth images captured for the HandID dataset. The corresponding colored image is included for better visualization.

This can be seen by their lower average per-joint-offset to the biased annotation (Avg Offset). We emphasize that these results are randomly selected and are representative of the whole HIM dataset.

References

[1] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a feedback loop for hand pose estimation. In *ICCV*, ICCV '15, pages 3316–3324, Washington, DC, USA, 2015. IEEE Computer Society.

[2] C. Wan, T. Probst, L. Van Gool, and A. Yao. Dense 3d regression for hand pose estimation. In *CVPR*, June 2018.

[3] X. Wu, D. Finnegan, E. O'Neill, and Y.-L. Yang. Handmap: Robust hand pose estimation via intermediate dense guidance map supervision. In *ECCV*, September 2018.

[4] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim. Bighand2.2m benchmark: Hand pose dataset and state of the art analysis. In *CVPR*, July 2017.

[5] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei. Model-based deep hand pose estimation. In *IJCAI*, IJCAI'16, pages 2421–2427. AAAI Press, 2016.

Visualization of Predictions on the Biased HIM dataset

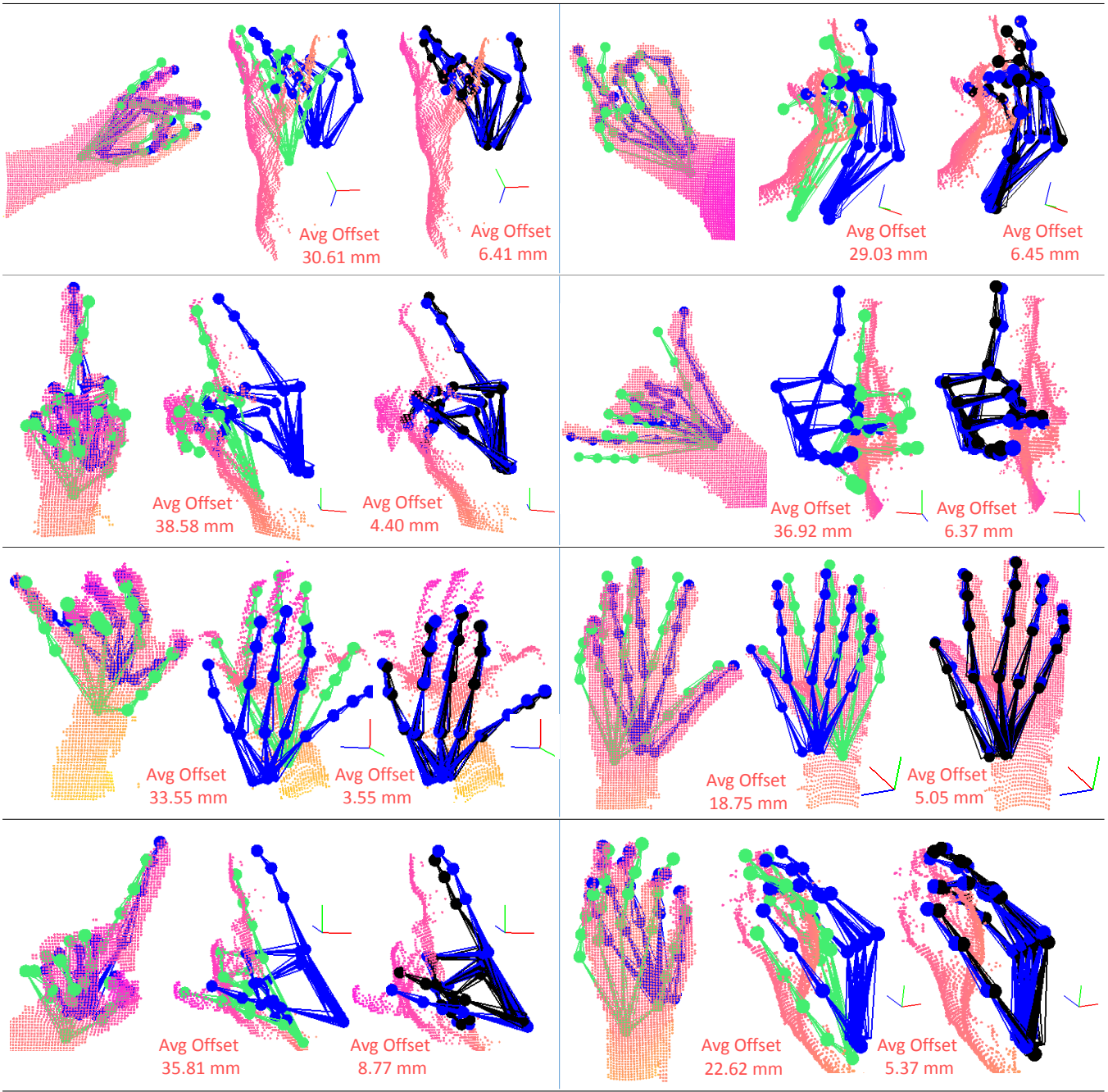

Fig. 4: Our predictions are in **green**, the "ground truth" annotations are in **blue**, and example State-of-the-Art (SotA) [3] predictions are in **black**. Each cell shows the camera view (**left**), a novel view (**middle**), and the same novel view with SotA predictions (**right**).