SemiHand: Semi-supervised Hand Pose Estimation with Consistency

Linlin Yang[1,2], Shicheng Chen[1], Angela Yao[1]
[1]National University of Singapore, Singapore
[2]University of Bonn, Germany

Abstract

We present SemiHand, a semi-supervised framework for 3D hand pose estimation from monocular images. We pretrain the model on labelled synthetic data and fine-tune it on unlabelled real-world data by pseudo-labeling with consistency training. By design, we introduce data augmentation of differing difficulties, consistency regularizer, label correction and sample selection for RGB-based 3D hand pose estimation. In particular, by approximating the hand masks from hand poses, we propose a cross-modal consistency and leverage semantic predictions to guide the predicted poses. Meanwhile, we introduce pose registration as label correction to guarantee the biomechanical feasibility of hand bone lengths. Experiments show that our method achieves a favorable improvement on real-world datasets after fine-tuning.

1. Introduction

A key challenge of monocular 3D hand pose estimation is getting sufficient high-quality ground-truth poses. Labelling real-world data to an accurate enough degree often requires dedicated interfaces and or multi-view camera rigs. This makes it non-trivial to gather “in-the-wild” data that is much sought-after for actual application deployment.

Synthesizing training data is considered an easy alternative to get accurate labels and has been incorporated into many learning-based frameworks. Yet there exists a significant domain gap between synthetic and real-world images so the performance of models trained on synthetic data deteriorates significantly when applied to real-world data. The favoured approach for reducing the domain gap is a mix-and-train strategy [12], *i.e.* mixing multiple real-world datasets together with synthetic data for training. Such a strategy depends largely however on the quantity and quality of the labelled samples in the combined datasets.

What if we tried to learn only from labelled synthetic data and fully unlabelled real-world data? We target exactly this scenario and present the first framework for domain-separated semi-supervised learning for 3D hand pose estimation. A classic approach in semi-supervised learning is to generate pseudo-labels [16] for the unlabelled data, usually via a classifier learned from the labelled portion of the data [16, 25]. The utility of pseudo-labels is highly variable. Used naively, these labels are even detrimental to learning because of confirmation bias [1], *i.e.* , the classifier over-fits to the pseudo-labels which tend to be noisy and or inaccurate, so additional corrections are necessary [1, 11, 43, 38]. Additionally, consistency training with unlabelled data [25, 1, 34] can increase the reliability of pseudo-labels.

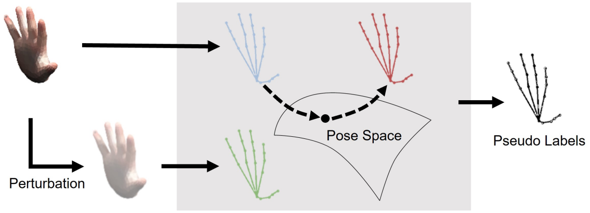

Figure 1: Pseudo-labelling of SemiHand. Our pseudo-label with confidence is generated based on the prediction from original (blue pose), the prediction from perturbation (green pose) and the corrected prediction (red pose).

We integrate these concepts and introduce SemiHand, a framework that considers spatial consistency and biomechanical feasibility for semi-supervised hand pose estimation. We propose two consistency losses to encourage the predictions to be consistent with perturbations and other modalities. As our labelled and unlabelled data come from different domains, *i.e.* synthetic vs real RGB images, there is the added challenge of domain adaptation to the unlabelled data. To bridge the domain gap, we propose a cross-modal consistency and leverage semantic predictions [19] from an auxiliary task to provide guidance for the predicted poses. Meanwhile, we regard predictions on real-world data as noisy labels; further training the network from these predictions directly may actually be detrimental due to their inaccuracy. To mitigate the impact of this confirmation bias, we introduce label correction and sample selection based on the feasibility so that we train with only corrected pseudo-labels with high-confidence. We show our pseudo-labelling strategy in Fig. 1.

Pseudo-labelling and consistency training are already established in semi-supervised classification [16, 25, 1]. However, extending such concepts for a regression task and in the context of 3D pose estimation is non-trivial and we are the first to present a unified framework to do so. For example, existing methods [11, 32] primarily learn a noise transition matrix to correct pseudo-labels; such an approach is not applicable for regression and we instead focus on the confidence and feasibility of poses as a selection and correction criteria. Similarly, consistency training in classification simply keeps the predicted categories unchanged under perturbation. Consistency in 3D pose estimation however needs to account for the change in label, *i.e.* the pose after perturbation. We summarize our contributions below:

- We propose a novel RGB-based hand pose estimation framework using labelled synthetic data and unlabelled real-world data; it is the first semi-supervised framework that combines pseudo-labeling with consistency training for RGB-based hand pose.
- Based on the feasibility of hand poses, we propose a method for pose registration and sample selection to correct noisy label outputs and select pseudo-labels of high confidence for training.
- We propose two consistency losses for 3D pose estimation to encourage the predictions to be consistent with perturbations and auxiliary modalities.
- Using a pre-trained synthetic model, we are able to adapt our model to challenging real-world datasets without any labels. Our results are compelling when compared to fully supervised frameworks and outperform previous works on synthetic image enhancement.

2. Related work

2.1. 3D Hand Pose Estimation

Most recent methods apply deep learning and propose dedicated network architectures and or training strategies, *e.g.* voxel-to-voxel predictions [20], point-to-point regression [10, 17], and pixel-wise estimations [8, 12]. Other works like [7] propose a tree-like network structure to capture the hand's topology. As for training strategies, existing works are diverse and have explored multitask learning [5, 4, 41], multi-view constraints for self-supervision [30, 31], and biomechanical constraints [26] as regularization. In RGB-based hand pose estimation, datasets are still relatively small and highly variable from each other. As such, most approaches cannot generalize to other datasets or in-the-wild scenarios. To improve the cross-dataset generalization, existing works like [12] adopt a mix-and-train strategy, *i.e.* , mix multiple real-world datasets together with synthetic data for training. Following this approach, most RGB-based works tend to synthesize more training samples using a GAN [21] or a generation model [14].

For 2D pose, semi-supervised learning methods like [23] treat each 2D keypoint independently and select 'labels' based on heatmap peaks. For 3D pose, weakly- and semi-supervised learning explore using weak labels or simply unlabelled data to improve cross dataset performance. Works like [4, 2] use 2D pose or the hand mask as weak labels while projecting the points in 3D to image coordinates.

Self-supervised learning for 3D pose removes even the requirement of weak labels, The most related works are for depth-based inputs [6, 30, 31] and human pose estimation [13]. Depth-based works like [6] use point cloud reconstruction as an auxiliary task to improve the performance of 3D hand pose estimation. Beyond that, Wan *et al.* [30, 31] introduce model-fitting with differentiable renderers for depth map reconstruction to utilize unlabelled data. RGB images however are affected by illumination and complex backgrounds, which prevent direct application of reconstruction or rendering approaches to RGB. As for the RGB-based human pose estimation, existing work [13] focuses on unlabelled multi-view images, which is still a highly limited scenario.

2.2. Semi-supervised learning

Consistency training and pseudo-labeling has recently shown much promise for semi-supervised classification [25, 34, 11, 3, 1, 29] and segmentation [43, 9]. Recent semi-supervised works have achieved comparable performance to supervised methods with only a fraction of the labels. For consistency training, works like [34, 9] have explored various augmentations. The mean teacher strategy [29] accelerates consistency training by averaging model weights instead of label predictions. For pseudo-labeling, operations such as argmax [16], sharpening [3] or thresholding [25] have been introduced to modify predictions as labels. Others [1, 11, 43, 38] treat predictions as noisy labels and introduce label correction to generate pseudo-labels.

Our work is the first to explore pseudo-labelling and consistency learning for hand pose estimation. Several distinctions separate pose estimation from the previous application of these techniques for image classification and segmentation. Formulation-wise, it is a regression problem that critically depends on spatial information. Secondly, there is a clear separation between biomechanical feasible versus infeasible poses. Therefore, we design a novel pipeline for semi-supervised hand pose estimation with corrected pseudo-labels and spatial consistency.

3. Methodology

We present an overview of our framework in Fig. 2. For pose estimation, let $X_L = \{(\mathbf{x}_i^l, \mathbf{p}_i, \mathbf{w}_i) : i \in (1, \cdots, N)\}$

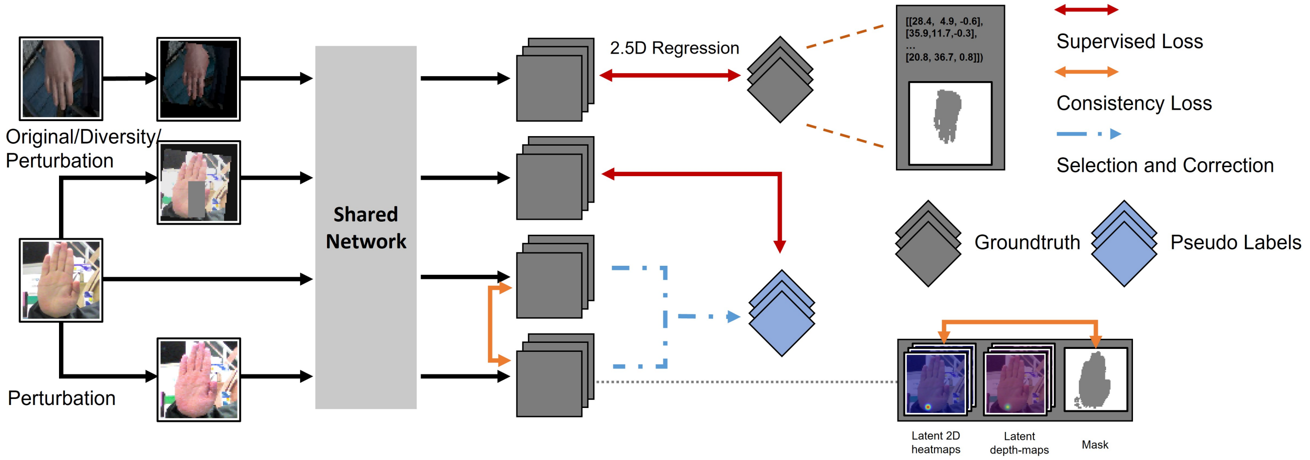

Figure 2: Overview of SemiHand. The model is pre-trained on labelled synthetic data. Consistency training (orange double headed arrow, see Sec. 3.3) on unlabelled real-world data with perturbation augmentations (see Sec. 3.4) and label correction and sample selection (blue dash-dotted arrow, See Fig. 1 and Sec. 3.2) together with augmentation of differing difficulties. (see Sec. 3.4).

be N labelled examples, where $\mathbf{x}_i^l$ is a labelled synthetic RGB image of a hand, $\mathbf{p}_i = (\mathbf{uv}_i, \mathbf{d}_i)$ is its target 2.5D hand pose, where $\mathbf{uv}$ is the the image pixel coordinates and $\mathbf{d}$ is its metric depth relative to the root keypoint, and $\mathbf{w}_i$ is a binary mask outlining the overall hand shape. Let $X_U = \{(\mathbf{x}_j^u) : j \in (1, \cdots, M)\}$ be M unlabelled examples, where $\mathbf{x}_i^u$ is an unlabelled real-world RGB image of a hand. We aim to estimate the 2.5D hand pose and its associated hand mask by learning a mapping f in the form of a neural network parameterized by θ, such that $(\mathbf{p}, \mathbf{w}) = f(\mathbf{p}, \mathbf{w}|\theta; X_L, X_U)$. In practice, the hand mask $\mathbf{w}$ is obtained by our shared fully convolutional network though our formulation is sufficiently general that it can also be learned by a separate network. We optimize a mixed objective of

$$\mathcal{L} = \mathcal{L}_{\text{sup}}(X_L) + L_{\text{unsup}}(X_U) + \lambda_c \mathcal{L}_{\text{cons}}(X_L, X_U), \quad (1)$$

where $\mathcal{L}_{\text{sup}}$ is the supervised loss, $\mathcal{L}_{\text{unsup}}(X_U)$ is an unsupervised loss with pseudo-labels and $\mathcal{L}_{\text{cons}}(X_L, X_U)$ is a consistency loss. λ_c is a hyperparameter. In the following, we introduce the details of the three losses.

3.1. Supervised Pose Estimation

A standard approach for 3D hand pose estimation is 2.5D pose regression [12] followed by a lifting into full 3D if camera intrinsics are known. The main benefit of regressing pose in 2.5D is the pixel-wise representation. This adds flexibility for multitask learning and can easily be extended to predict other pixel-wise outputs such as segmentations or depth maps with fully convolutional networks. The multitasking strategy achieves improvement for hand pose estimation [35]. In our work, besides 2.5D pose $\mathbf{p}$, we also predict hand mask $\mathbf{w}$. We show the details of 2.5D regression in the supplementary material. Here, we first define the distance ℓ between two 2.5D poses $\mathbf{p}_1 = (\mathbf{uv}_1, \mathbf{d}_1)$ and $\mathbf{p}_2 = (\mathbf{uv}_2, \mathbf{d}_2)$ as

$$\ell(\mathbf{p}_1, \mathbf{p}_2) = ||\mathbf{uv}_1 - \mathbf{uv}_2||_2^2 + \lambda_{\mathbf{d}} ||\mathbf{d}_1 - \mathbf{d}_2||_2^2, \quad (2)$$

where $\lambda_{\mathbf{d}}$ is a hyperparameter with a value of 50 in our paper. Given a ground-truth $\mathbf{p}_{gt}$, $\mathbf{w}_{gt}$ and corresponding predictions $\mathbf{p}$, $\mathbf{w}$, the supervised loss is defined as:

$$\mathcal{L}_{\text{sup}}(X_L) = \ell(\mathbf{p}, \mathbf{p}_{gt}) + \lambda_{\mathbf{w}} ||\mathbf{w} - \mathbf{w}_{gt}||_1, \quad (3)$$

where $\lambda_{\mathbf{w}}$ is a hyperparameter. In this paper, we adopt the two-stacked hourglass with 2.5D regression as our backbone to estimate 2.5D representation and hand mask.

3.2. Pseudo-labels for Pose Estimation

For now, assume we have some initial network $f(\theta)$ from pre-training. We initialize pseudo-labels $\hat{\mathbf{p}} = (\hat{\mathbf{uv}}, \hat{\mathbf{d}})$ of X_U using the prediction of $f(\theta)$ and fine-tune the model with corrected pseudo-labels $\mathbf{r}$. With the prediction $\mathbf{p}$ from $f(\mathbf{p}|\theta; X_U)$, the objective $\mathcal{L}_{\text{unsup}}(X_U)$ can be formulated as:

$$\mathcal{L}_{\text{unsup}}(X_U) = \mathbb{1}(\mathcal{C}(\hat{\mathbf{p}}) \leq \tau)\ell(\mathbf{p}, \hat{\mathbf{p}}), \quad \text{where } \hat{\mathbf{p}} \backsim \mathcal{M}. \quad (4)$$

Here, $\mathbb{1}(\cdot)$ is the indicator function, $\mathcal{C}(\cdot)$ is a function to estimate the confidence of given pseudo-labels, and τ is a confidence threshold. Pseudo-labels are often noisy and may require corrections [18, 11]. In this objective, we constrain the pseudo-pose $\hat{\mathbf{p}}$ to be drawn from $\mathcal{M}$, a pose space whose points are biomechanical feasible poses in which bone lengths are consistent with the given hand model. Based on Eq. 4, we introduce a pose registration function

$P(\cdot)$ to project the pseudo-labels $\hat{\mathbf{p}}$ to corrected poses $\mathbf{r}$ and add a loss to minimize the distance between the prediction $\mathbf{p}$ and $\mathbf{r}$. To prevent degenerate labels $\mathbf{r}$, we add a regularizer to encourage $\mathbf{r}$ to remain close to $\hat{\mathbf{p}}$. Adding these terms, we get

$$\mathcal{L}_{\text{unsup}}(X_U) = \mathbb{1}(\mathcal{C}(\hat{\mathbf{p}}) \leq \tau)\ell(\mathbf{p},\hat{\mathbf{p}}) + \ell(\mathbf{r},\mathbf{p}) + \ell(\mathbf{r},\hat{\mathbf{p}}), \quad (5)$$

with $\mathbf{r} = P(\hat{\mathbf{p}})$. For learning the network θ and the pseudo-labels $\hat{\mathbf{p}}$, We solve the objective iteratively. First, we update the parameter of the network θ by

$$\mathcal{L}_{\text{unsup}}(X_U) = \mathbb{1}(\mathcal{C}(\hat{\mathbf{p}}) \leq \tau)\ell(\mathbf{p},\hat{\mathbf{p}}) + \ell(\mathbf{r},\mathbf{p}), \quad (6)$$

which can be solved by gradient descent. We then estimate the pseudo-labels $\hat{\mathbf{p}}$ and its correction $\mathbf{r}$ based on the previous prediction $\mathbf{p}'$ and the previous correction $\mathbf{r}'$,

$$\begin{aligned} \hat{\mathbf{p}} &= \arg\min_{\hat{\mathbf{p}}} \ell(\mathbf{p}',\hat{\mathbf{p}}) + \ell(\mathbf{r}',\hat{\mathbf{p}}), \\ \mathbf{r} &= P(\hat{\mathbf{p}}). \end{aligned} \quad (7)$$

Label Correction. Estimating the joint locations independently is not effective to ensure the biomechanical feasibility of the hand. Inspired by the similarity transformation of [31], we propose a pose registration function P. More specifically, we estimate the transformation T with a greedy approximation based on the hand's kinematic chain. As shown in Fig. 4 right, given a template (black) and a prediction (gray), we first align the root by translation, and then calculate the bone direction (dotted gray line) using the parent node of registered pose and the child node of estimation. With calculating T of each bone along with the chain of a hand, we get the registered pose (orange). The proposed greedy approximation avoids the accumulation of end point errors and ensure the feasibility of bone lengths without any training. More details are provided in the supplementary.

Sample Selection. We design the confidence function $\mathcal{C}$ for samples based on the plausibility and stability of the pseudo-labels $\hat{\mathbf{p}}$ for the unlabelled data $\mathbf{x}^u$ as below:

$$\mathcal{C}(\hat{\mathbf{p}}) = \ell(\mathcal{T}(\hat{\mathbf{p}}), f(\mathbf{p}|\theta;\mathcal{T}(\mathbf{x}^u))) + \ell(\hat{\mathbf{p}}, P(\hat{\mathbf{p}})), \quad (8)$$

where $\mathcal{T}$ is a random perturbation augmentation. The proposed confidence is a sum of the distance between the prediction of perturbed image and its corresponding pseudo-label, and the distance between the pseudo-label and its corrected pseudo-label.

3.3. Self Consistency for Pose Estimation

Algorithm 1 Semi-supervised hand pose estimation.

```
Require: Pre-trained model $\theta_0$ based on $\mathcal{L}_{sup}$, threshold $\tau$,
    epoch number $K$, $X_L$ and $X_U$
Ensure: Final model $\theta$ and pseudo labels $\hat{\mathbf{p}}$
Initialize the pseudo-labels $\hat{\mathbf{p}}$ for $X_U$
Initialize the corrected pseudo-labels $\mathbf{r}$ for $X_U$
for $t = 1, \ldots, K$ epochs do
    Calculate $\mathcal{C}(\hat{\mathbf{y}})$
    Update $\theta$ via gradient ascent of Eq. 6 with $\mathcal{L}_{sup}(X_L)$
        and $\mathcal{L}_{cons}(X_L, X_U)$
    Update $\hat{\mathbf{p}}$ and $\mathbf{r}$ based on Eq. 7
end for
```

For both X_L and X_U, we introduce a view consistency term $\mathcal{L}_{\text{vc}}$ and a cross-modal consistency term $\mathcal{L}_{\text{cc}}$ to improve generalization. The consistency loss $\mathcal{L}_{\text{cons}}(\mathcal{X}_L, \mathcal{X}_U)$ is simply the sum of the two:

$$\mathcal{L}_{\text{cons}} = \mathcal{L}_{vc} + \mathcal{L}_{cc}. \quad (9)$$

View Consistency. As shown in Fig. 3, we augment the training samples by rotating or translating the samples, as depicted in Sec. 3.4, and encourage transformed 2.5D predictions to be consistent with predictions of the transformed samples like existing 2D works [23]. The proposed loss function, with random perturbation $\mathcal{T}$ is:

$$\begin{aligned} \mathcal{L}_{vc} = &\ \ell(f(\mathbf{p}|\theta;\mathcal{T}(\mathbf{x})), (\mathcal{T}(f(\mathbf{p}|\theta;\mathbf{x})))) \\ &+ ||f(\mathbf{w}|\theta;\mathcal{T}(\mathbf{x})) - (\mathcal{T}(f(\mathbf{w}|\theta;\mathbf{x})))||_1. \end{aligned} \quad (10)$$

This loss encourages more robust and stable predictions for unlabelled data X_U.

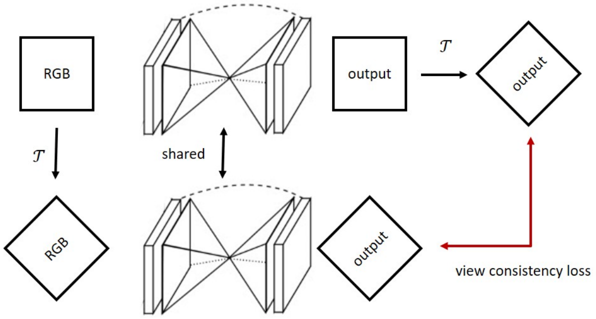

Figure 3: Overview of view consistency loss.

Cross-modal Consistency. Zamir *et al.* [39] observed that learning with cross-modal consistency improves prediction accuracy. In that regard, different modality representations *e.g.* RGB image, depth map, of the same hand should be 'consistent' in their pose. But how can we enforce this consistency across these modalities without actual pose labels? In this case, we incorporate multi-task learning and estimate multi-modal outputs *i.e.* pose and mask, and add a model-fitting energy term. The proposed energy function encourages consistency between the 2D pose and the hand mask, which we find improves pose and overall generalization. Additionally, we adopt a stop-gradient operation stop$(\cdot)$ to the mask as shown in Fig. 5 to prevent inaccurate poses from degenerating the masks.

Specifically, we approximate the hand mask with 55 circles: 9 for each finger and 10 for the palm. The circle hand model is parameterized as $\mathbf{m} = \{m^0, \cdots, m^{54}\}$, where $m^i = (c^i, r^i)$ is the i^{th} circle centered at c^i with radius r^i. The circle centers are manually defined based on the 2D pose, while radii are pre-trained from synthetic data. Fig. 4 middle shows an example of the approximated hand mask and the circles for the little finger.

The cross-modality consistency loss $\mathcal{L}_{\text{cc}}$ is the sum of two standard model-fitting energy terms:

$$\mathcal{L}_{\text{cc}}(\mathbf{uv}, \mathbf{w}) = \mathcal{L}_{\text{m2d}}(\mathbf{uv}, \text{stop}(\mathbf{w})) + \mathcal{L}_{\text{d2m}}(\mathbf{uv}, \text{stop}(\mathbf{w})). \tag{11}$$

The model-to-data term $\mathcal{L}_{\text{m2d}}$ is an L1 distance encouraging the circle-approximated mask to be as similar as possible to the estimated mask:

$$\mathcal{L}_{m2d}(\mathbf{uv}, \mathbf{w}) = ||R(G(\mathbf{uv})) - \mathbf{w}||_1, \tag{12}$$

where $G(\cdot)$ estimates the centers and radius based on the 2D hand pose and $R(\cdot)$ renders the circles to a hand mask like [30]. Note that this term has no gradients on the background of the rendered mask. Hence, we add a data-to-model term $\mathcal{L}_{d2m}$ to measure the registration error between the estimated hand model and hand mask:

$$\mathcal{L}_{d2m}(\mathbf{uv}, \mathbf{w}) = \sum_{g \in \Omega} d(\mathbf{w}(g), G(\mathbf{uv})), \tag{13}$$

where Ω is the set of all pixel locations and the distance function $d(\cdot)$ is defined as:

$$d(\mathbf{w}(g), \mathbf{m}) = \begin{cases} \max(\min\limits_{i \in [0,54]}(||g - c^i||_2 - r^i), 0) & \text{if } \mathbf{w}(g) = 1, \\ \max(\max\limits_{i \in [0,54]}(r^i - ||g - c^i||_2), 0) & \text{otherwise.} \end{cases} \tag{14}$$

Specifically, the distance estimates pixel g's distance to the nearest circle m^i with radius r^i centered at c^i. If the predicted mask value at g is correct, the distance is set to 0. More details on the consistency loss can be found in the supplementary material.

3.4. Data Augmentation

Initially, we found that adding view-point consistency to be non-convergent. We speculated the cause to be mode collapse, *i.e.* all the 2D pose predictions gradually move to the center of the image. A similar phenomenon was observed in FixMatch [25]; they found that data augmentation of differing difficulties could improve training stability. As such, we also adopt two types of data augmentation like [25], as shown in Fig. 2. Specifically, we introduce diversity augmentation for the labelled and high-confidence pseudo-labelled data and perturbation augmentation for unlabelled data respectively, which we found to mitigate the problem of mode collapse.

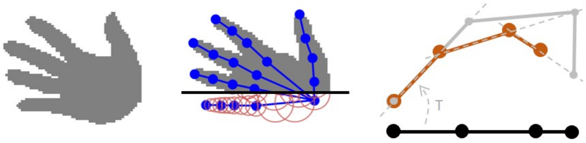

Figure 4: Hand model and pose registration. Left: the ground-truth hand mask; Middle: Our rendered hand mask based on ground-truth 2D pose (blue points); Right: pose registration of the template hand (black) to observed joints (grey) to result in a registered hand (orange). Figure best viewed in colour.

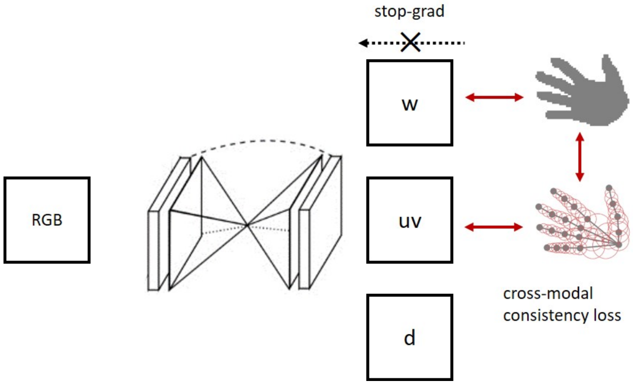

Figure 5: Overview of cross-modal consistency loss. (uv, d) are 2.5D hand outputs; $\mathbf{w}$ *denotes the hand mask.*

In all of our experiments, diversity augmentation is similar to augmentations used in existing supervised learning methods [12, 5, 36]. It includes color jitter, translation, rotation, scale, gray-scale and random erasure. Differently, for unlabelled data, we simply perturb with translations of [-5,5] pixels or rotations of either $[-2°, 2°]$ or 90°,180° and 270°.

4. Experiments

4.1. Implementation Details

In the experiments, we adopt the two-stacked hourglass as our backbone. The input and output resolution are both 64×64. We set the hyperparameters from Eqs. 1 to 4 empirically, with $\lambda_c = 0.1$, $\lambda_{\mathbf{d}} = 50$, $\lambda_{\mathbf{w}} = 100$ and $\tau = 1.5$. For pre-training on the synthetic data, we use an Adam optimizer with an initial learning rate of 10^{-3} and a batch size of 32. We train the model for 100 epochs, lowering the learning rate by a factor of 10 at the 60th and 90th epoch. For fine-tuning, we use the learning rate 10^{-4} and a batch size of 128. We set K to 10. At 5th iteration, we lower the learning rate to 10^{-5}. The associated algorithm is shown in Alg. 1.

4.2. Datasets and Evaluation metrics

Our method is trained on one synthetic dataset, the Rendered Hand Pose Dataset (RHD) [44] and evaluated on four real-world datasets, Stereo Hand Pose Tracking Benchmark (STB) [40], Dexter+Object Dataset (DO) [28], Hand-3D-Studio (H3D) [42] and YouTube 3D Hands (YT3D) [15].

To further verify the effectiveness of our proposed method, we also introduce and evaluate on a new real-world hand sequence dataset (HSD). HSD is a video dataset with 3D poses annotated in a semi-automated fashion like [45]. It consists of 4 sequences. Each sequence is performed by one actor and contains 20K frames. We use the first two sequences for training and others for testing. More details of this dataset can be found in the supplementary.

To evaluate the accuracy of estimated poses, we use two common metrics: (1) mean end-point-error (EPE), measuring the average Euclidean distance between predicted and ground-truth joints, and (2) area under the curve (AUC) on the percentage of correct keypoints (PCK) curve based on certain error thresholds. For a fair comparison with state-of-the-art, we follow [27, 36], assuming that the global hand scale and the hand root position are known, and set the middle finger's base position as the hand root. For convenience, we also assume that hand template is given. For H3D and YT3D, we use 40 mm from STB as reference bone length defined by [45]. Our default setting is fine-tuning with only the training data of a (single) real-world dataset's training partition. Following the convention of [30], the test data is withheld completely. Additionally, we use the labels of these real-world datasets only for evaluation purposes.

4.3. Ablation Study

Baseline. To start with, we first investigate the domain gap that exists between the synthetic RHD versus the real-world STB. The pre-trained network, trained and tested on RHD achieves good performance with a mean EPE 12.08 mm. However, the same network's errors almost double to a mean EPE of 23.41 mm and 23.83 mm on the STB training and testing datasets respectively (see 'baseline' method in Tab. 1). If we train the network only on STB, it is prone to over-fitting due to the small size of the dataset, so it leads to a large error on testing data (18.04 mm). If one merges the training datasets of RHD and STB in a mix-and-train strategy, we can lower this error to 7.32 mm and this serves as the upper bound in performance for semi-supervised methods.

Method	training set	testing set	
		STB train	STB test
baseline	RHD train(w/)	23.41	23.83
baseline	STB train(w/)	5.27	18.04
baseline	RHD train(w/) STB train(w/)	5.25	7.32
with vc	RHD train(w/) STB train(w/o)	19.98	21.03
with cc		20.59	20.92
with vc+cc		19.18	19.93
with pseudo-labeling	RHD train(w/) STB train(w/o)	15.68	16.31
our proposed	RHD train(w/) STB train(w/o)	13.82	14.60
our proposed	RHD train(w/) STB test(w/o)	15.83	14.51
our proposed	RHD train(w/) STB train+test(w/o)	13.78	13.95

Table 1: Ablation study with mean EPE [mm]. w/ and w/o indicates with and without labels for training.

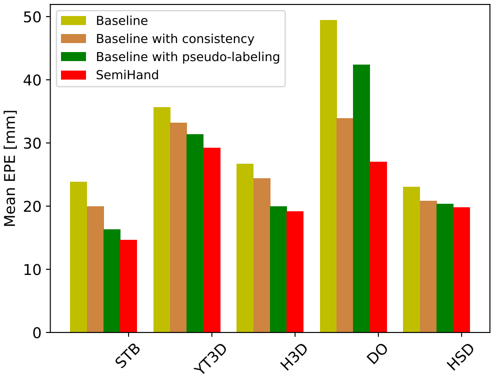

Figure 6: Comparison of baseline, with only consistency training, with only pseudo-labeling and our proposed SemiHand. Our proposed two modules both improve the performance with respect to the baselines, and their combination further leads to a higher accuracy.

Impact of our components. We next analyse the performance of our method's individual components to isolate the impact of consistency training and pseudo-labeling. We fine-tune the pre-trained model with only view consistency loss (with vc), only cross-model consistency loss (with cc), both consistency losses (with vc+cc) and with pseudo-labelling in Tab. 1. Each component improves the performance; adding pseudo-labelling achieves an impressive 7.52 mm improvement on the STB testing set. Combining these components further decreases the error. With both consistency training and pseudo-labeling, we achieve a 9.23mm improvement on the STB testing set with fine-tuning on the unlabelled STB training set.

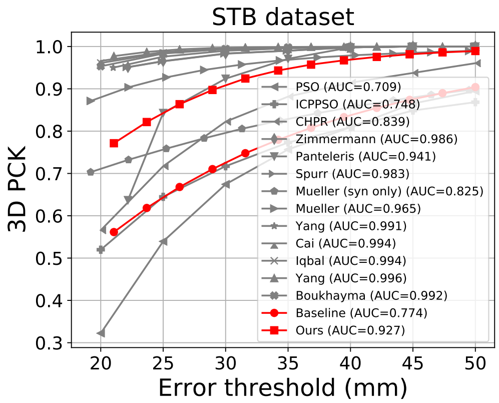

Figure 7: AUC: Comparison to state-of-the-art on STB. Our SemiHand improves the baseline's AUC and achieves comparable performance to other supervised learning methods.

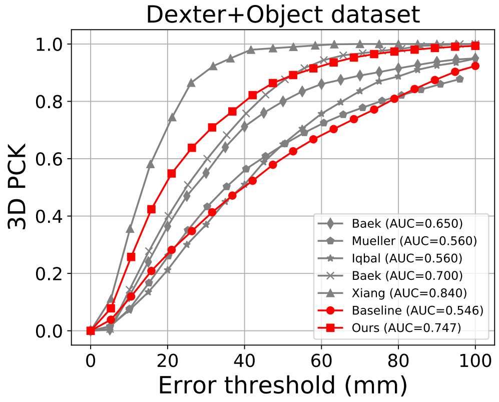

Figure 8: AUC: Comparison to state-of-the-art on DO. Our SemiHand improves the baseline's AUC and outperforms some supervised learning methods using the mix-and-train strategy.

For further verification, we compare the following: (1) baseline, (2) baseline with consistency training, (3) baseline with pseudo-labels and (4) our proposed method on all real-world datasets (see results in Fig. 6). We can see that both consistency training and pseudo-labeling can improve the performance with respect to the baselines. Furthermore, the combination of our two modules leads to a higher accuracy. With our semi-supervised fine-tuning, we achieve a decrease in mean EPE of up to 9.2 mm on STB, 22.4 mm on DO, 6.4 mm on YT3D, 7.46 mm on H3D and 3.3 mm on HSD as shown in Fig. 6. The full model is comparable to existing supervised methods.

Impact of training data. In Tab. 1 under 'our proposed', we fine-tune the network on different STB sets, *i.e.* , STB train set only, STB test set only and both. We find that fine-tuning on the testing image directly achieve lower mean EPE (13.82 mm/13.78 mm versus 15.83 mm for STB train and 14.51 mm/13.95 mm versus 14.60 mm for STB test). Moreover, as the amount of unlabelled training data increases, the mean EPE decreases correspondingly. As shown in Tab. 1, fine-tuning with both STB train and test sets outperforms fine-tuning independently. We also verify this by fine-tuning with different percentages of STB training data in Fig. 10. We decrease the mean EPE of STB test set from 17.31mm to 14.60 mm by increasing the percentage of unlabelled STB training data during training.

4.4. Comparison to state-of-the-art

We compare our hand pose estimation results with state-of-the-art methods [2, 21, 12, 33, 37, 36, 4, 27, 22], on STB and DO as shown in Fig. 7 and 8. We can see that after fine-tuning, our SemiHand improves the baseline's AUC significantly (0.774 to 0.927 for STB, 0.546 to 0.747 for DO). For STB, our semi-supervised method achieves comparable performance to other supervised learning methods, even without any labels of STB. The work [21] also reports its performance training on synthetic data only. As shown in Fig. 7, ours outperforms [21] by a large margin (0.927 vs. 0.825).

Many existing methods use DO to evaluate cross-dataset performance. Our proposed semi-supervised method outperforms most existing supervised methods, even though they mix-and-train RHD with other synthetic data [2, 21], STB [40], MPII+NZSL [24] or MVBS [24]. This confirms our original motivation of exploiting unlabelled RGB images and improving the accuracy of pose estimation. Note that [33] does report better performance but they incorporate a large-scale (111K) labelled real-world dataset for training.

With our proposed semi-supervised method, the predictions of unlabelled data will gradually converge. We show two qualitative examples of the gradual convergence from the predictions of pre-trained model to our stable predictions in Fig. 9. Interestingly, we also find cases like the example shown in Fig. 9, where our predictions seem more accurate than the manually annotated ground-truth, *i.e.* predicted keypoints are centered on the finger, while labelled keypoints lie at the edge of the fingers. Given the saturated results of state-of-the-art methods on STB, it is likely that many networks are over-fitting to manual annotation biases or noise.

4.5. Comparison to weakly-supervised methods

As our SemiHand is the first semi-supervision framework for 3D hand pose estimation from monocular images, there are no direct comparable methods. We compare instead to a weakly-supervised method [4]. We fine-tune the pre-trained model on m% STB training data, either without any labels (ours, SemiHand), with ground-truth (strong supervision) and with weak labels of either 2D poses or masks. The percentage of STB training set is varied from

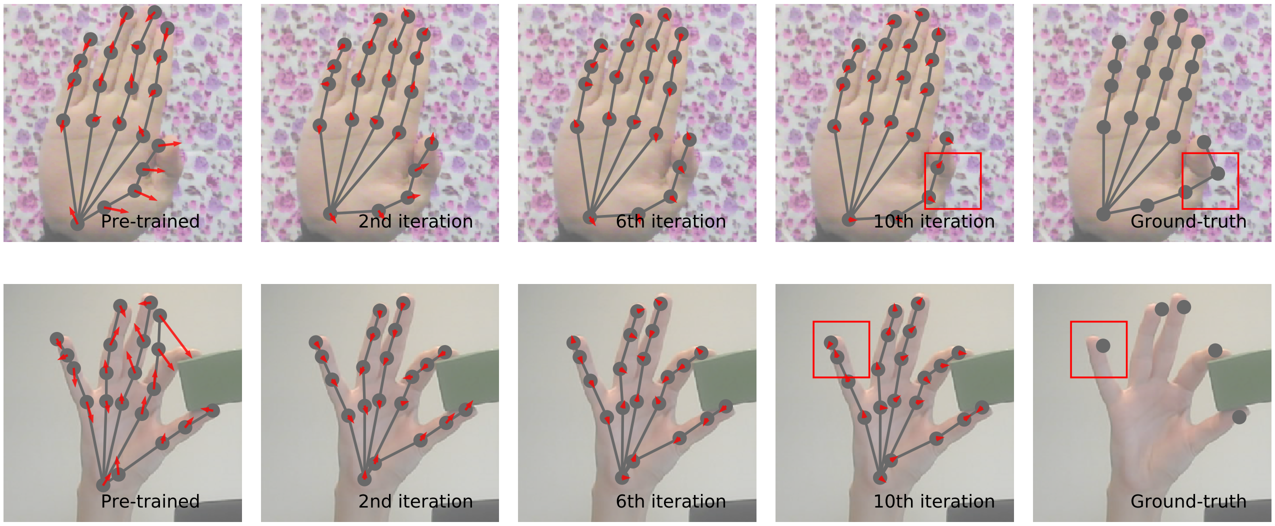

Figure 9: Gradual convergence from the prediction of pre-trained model to our final prediction. The arrows indicate the direction and distance of prediction movement during fine-tuning. For 10thiteration, the optimization converges because the length of arrows become almost zeros. We highlight the differences between our stable predictions and the ground-truth poses with red boxes. Figure best viewed in colour.

5% to 100% to compare the mean EPE on STB testing set. As shown in Fig. 10, when fine-tuning with masks or 2D poses as weak labels, the weakly-supervised method [4] achieves 4.0 mm and 7.1 mm improvement on STB testing set respectively. This indicates that 2D pose provides stronger supervision than simply a mask. Meanwhile, without any labels, our SemiHand achieves a 9.2 mm improvement, demonstrating the effectiveness of our method compared to [4]. Note that we discuss only the relative improvement as we use a different backbone than [4]. Given that adding even a small amount of labels (as per the fully supervised method) is still better, this encourages us to further explore the use of unlabelled images.

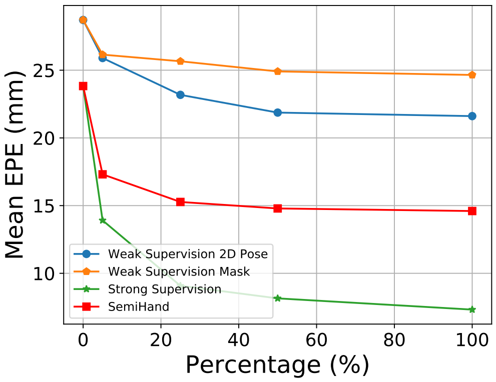

Figure 10: Mean EPE on STB testing data with fine-tuning on different percentage of STB training data. As the amount of training data increases, SemiHand achieves a similar trend as the weakly-supervised methods, i.e., the mean EPE decreases correspondingly.

5. Conclusions

We aim to develop a semi-supervised 3D pose estimation framework, using labelled synthetic and unlabelled real-world data. Directly applying the existing semi-supervised method is nontrivial because pose estimation is a regression problem that critically depends on spatial information. We therefore designed a new framework based the pose feasibility and spatial consistency, with pseudo-labels and consistency training. Experiments on different datasets demonstrate that our approach successfully leverages real-world RGB images without any labels, paving a path forwards for learning pose estimation systems with only synthetic labels. In the future, we would like to explore domain adaptation methods and more consistencies over time and or multiple views to further improve the accuracy. Also, we will explore different frameworks like the teacher-student framework or the Siamese framework for semi-supervised pose estimation.

Acknowledgments This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2020-016). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, pages 1–8. IEEE, 2020. 1, 2

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *CVPR*, pages 1067–1076, 2019. 2, 7

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 2

[4] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019. 2, 7, 8

[5] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*, pages 666–682, 2018. 2, 5

[6] Yujin Chen, Zhigang Tu, Liuhao Ge, Dejun Zhang, Ruizhi Chen, and Junsong Yuan. So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning. In *ICCV*, pages 6961–6970, 2019. 2

[7] Kuo Du, Xiangbo Lin, Yi Sun, and Xiaohong Ma. Crossinfonet: Multi-task information sharing based hand pose estimation. In *CVPR*, pages 9896–9905, 2019. 2

[8] Linpu Fang, Xingyan Liu, Li Liu, Hang Xu, and Wenxiong Kang. Jgr-p2o: Joint graph reasoning based pixel-to-offset prediction network for 3d hand pose estimation from a single depth image. *arXiv preprint arXiv:2007.04646*, 2020. 2

[9] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*, 2020. 2

[10] Liuhao Ge, Zhou Ren, and Junsong Yuan. Point-to-point regression pointnet for 3d hand pose estimation. In *ECCV*, pages 475–491, 2018. 2

[11] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *ICCV*, pages 5138–5147, 2019. 1, 2, 3

[12] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, pages 118–134, 2018. 1, 2, 3, 5, 7

[13] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *CVPR*, pages 5243–5252, 2020. 2

[14] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, pages 4990–5000, 2020. 2

[15] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, June 2020. 6

[16] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013. 1, 2

[17] Shile Li and Dongheui Lee. Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. In *CVPR*, pages 11927–11936, 2019. 2

[18] Zhiwei Liu, Xiangyu Zhu, Guosheng Hu, Haiyun Guo, Ming Tang, Zhen Lei, Neil M Robertson, and Jinqiao Wang. Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In *CVPR*, pages 3467–3476, 2019. 3

[19] Adrian Lopez-Rodriguez and Krystian Mikolajczyk. Desc: Domain adaptation for depth estimation via semantic consistency. *arXiv preprint arXiv:2009.01579*, 2020. 1

[20] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *CVPR*, pages 5079–5088, 2018. 2

[21] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, pages 49–59, 2018. 2, 7

[22] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *WACV*, pages 436–445. IEEE, 2018. 7

[23] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, pages 4119–4128, 2018. 2, 4

[24] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, pages 1145–1153, 2017. 7

[25] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 1, 2, 5

[26] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. *arXiv preprint arXiv:2003.09282*, 2020. 2

[27] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, pages 89–98, 2018. 6, 7

[28] Srinath Sridhar, Franziska Mueller, Michael Zollhoefer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *ECCV*, 2016. 6

[29] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 2

[30] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Self-supervised 3d hand pose estimation through training by fitting. In *CVPR*, pages 10853–10862, 2019. 2, 5, 6

[31] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dual grid net: hand mesh vertex regression from single depth maps. In *ECCV*, pages 442–459. Springer, 2020. 2, 4

[32] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, pages 13726–13735, 2020. 2

[33] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, pages 10965–10974, 2019. 7

[34] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. 1, 2

[35] Lixin Yang, Jiasen Li, Wenqiang Xu, Yiqun Diao, and Cewu Lu. Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. *arXiv preprint arXiv:2008.05079*, 2020. 3

[36] Linlin Yang, Shile Li, Dongheui Lee, and Angela Yao. Aligning latent spaces for 3d hand pose estimation. In *ICCV*, pages 2335–2343, 2019. 5, 6, 7

[37] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *CVPR*, pages 9877–9886, 2019. 7

[38] Jiangchao Yao, Hao Wu, Ya Zhang, Ivor W Tsang, and Jun Sun. Safeguarded dynamic label regression for noisy supervision. In *AAAI*, pages 9103–9110, 2019. 1, 2

[39] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *CVPR*, pages 11197–11206, 2020. 4

[40] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016. 6, 7

[41] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, pages 2354–2364, 2019. 2

[42] Zhengyi Zhao, Tianyao Wang, Siyu Xia, and Yangang Wang. Hand-3d-studio: A new multi-view system for 3d hand reconstruction. In *ICASSP*, pages 2478–2482. IEEE, 2020. 6

[43] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*, pages 1–15, 2020. 1, 2

[44] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, pages 4903–4911, 2017. 6

[45] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, pages 813–822, 2019. 6