This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Cross-Domain 3D Hand Pose Estimation with Dual Modalities

Qiuxia Lin* Linlin Yang* Angela Yao
National University of Singapore
{qiuxia, yangll, ayao}@comp.nus.edu.sg

Abstract

Recent advances in hand pose estimation have shed light on utilizing synthetic data to train neural networks, which however inevitably hinders generalization to real-world data due to domain gaps. To solve this problem, we present a framework for cross-domain semi-supervised hand pose estimation and target the challenging scenario of learning models from labelled multi-modal synthetic data and unlabelled real-world data. To that end, we propose a dual-modality network that exploits synthetic RGB and synthetic depth images. For pre-training, our network uses multi-modal contrastive learning and attention-fused supervision to learn effective representations of the RGB images. We then integrate a novel self-distillation technique during fine-tuning to reduce pseudo-label noise. Experiments show that the proposed method significantly improves 3D hand pose estimation and 2D keypoint detection on benchmarks.

1. Introduction

Hand pose estimation supports a wide range of applications, including sign language recognition [18, 19] and gesture-based interaction systems [1]. However, it is difficult to obtain the large amounts of accurate ground truth labels required for training deep-learning-based hand pose estimation systems. Training models with synthetic data is one option, but such models exhibit a sim-to-real domain gap and generalize poorly to real-world settings. More sophisticated synthesis can narrow this gap, but the performance drop is still noticeable [23].

This paper addresses the cross-domain pose estimation problem and focuses on a semi-supervised setting. We target learning from labelled synthetic data and unlabelled real-world data for application to real-world data. Pre-training with synthetic data is common [14, 32]; surprisingly, only RGB synthetic images have been considered. Yet when generating synthetic data, it is relatively easy to render multiple data modalities. For example, the RHD dataset [41] features both RGB images and depth maps.

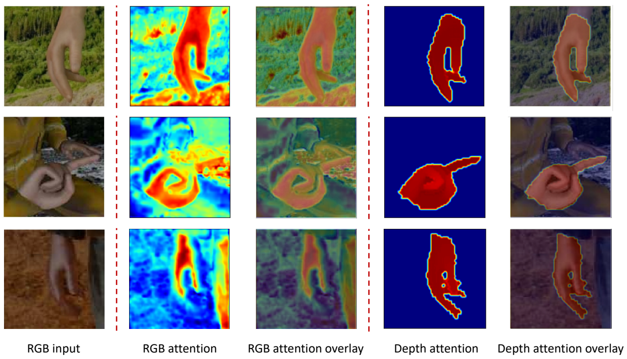


Figure 1. Attention comparisons between RGB and depth maps. Apart from the hand region, the RGB attention shows high activation on the background. The depth map attention however provides high activation only on the hand, confirming its strength to focus on task-relevant information.

Across the modalities, there are shared common visual cues relating to the underlying geometry or semantics that are task-relevant for hand pose estimation. To exploit such information during pre-training, a simple solution is to apply pixel-wise ℓ_1 alignment between feature maps of RGB and depth maps [37]. However, the large discrepancy between RGB images and depth maps might cause the pixel-wise alignment to also focus on irrelevant regions related to the background. Fig. 1 shows the attention derived from RGB versus depth maps and their overlays on the original RGB input. We observe that the RGB attention is strong in task-related areas, *i.e.* the hand, but also on unrelated areas of the background. In contrast, the attention from the depth map is successfully localized on the hand.

To focus more on relevant regions, we propose a dual-modality network for RGB images and depth maps that improves RGB-based hand pose estimation via a multi-level alignment. Specifically, we design an attention module to apply information learned from depth maps to the RGB image to produce attention-fused RGB features. The learned information is then aligned to RGB features by multi-modal supervision on the predictions from all modalities with ground truth. This limits the RGB encoder's sen-

*Equal contribution

sitivity to non-informative cues such as background or irrelevant textures [10, 34]. At the feature space level, we explore intra-modal and inter-modal contrastive learning for multi-modal data. Our work is the first to investigate supervised multi-modal contrastive learning on hand pose estimation. In particular, contrastive learning between RGB features and attention-fused RGB features minimizes feature discrepancies across modalities.

After pre-training, pseudo-labelling is commonly used to fine-tune on unlabelled data [4, 20, 22]. However, naïvely generated pseudo-labels are inevitably noisy and deteriorate model performance. To handle noisy pseudo-labels, we design a two-pass self-distillation procedure (see Fig. 2 (b)). The RGB input is first applied through the multimodal decoder to predict a depth map and pose. In a second pass, the same RGB input is applied together with the predicted depth map to estimate the pose. As the attention from depth maps activates RGB features in relevant regions, the two passes can be considered a refinement or denoising process. The pose predicted from the RGB in the first pass is encouraged to be consistent with the pose from the attention-fused version in the second pass. Such a procedure distills knowledge from within the network itself. The self-distillation also prevents the network from over-fitting to noisy samples that often have unstable predictions [17, 20].

In summary, we make the following contributions:

1. We propose a dual-modality network that learns from RGB images and depth maps but is applicable to only RGB inputs for fine-tuning and inference. The network features a specially designed attention module that identifies geometric relationships common to RGB and depth from stand-alone RGB images.

2. We propose the first supervised multi-modal contrastive learning method based on fused features to minimize feature discrepancies across modalities.

3. We introduce a self-distillation procedure to exploit yet not over-fit to noisy pseudo-labels during fine-tuning.

4. The proposed method significantly improves the state-of-the-art by up to **16.0%** and **14.8%** for 2D keypoint detection and 3D keypoint estimation respectively.

2. Related Work

2.1. Contrastive Learning

Contrastive learning encourages the learning of feature spaces in which similar sample pairs (positive pairs) stay close together while dissimilar samples (negative pairs) are further apart. It is applicable in unsupervised [6, 7, 28] and supervised [12, 15, 16] settings. Constructing beneficial positive-negative pairs forms the basis of contrastive learning. Existing works [6, 7, 16, 28] prefer to create positive pairs based on data augmentation. Interestingly, a recent work [28] introduced the use of different modalities of one instance as a positive pair, showing great potential. However, the large discrepancy between the different modalities may still limit the performance.

Pose estimation works [26, 40] use contrastive learning with unlabelled RGB images during pre-training. In contrast, we target labelled multi-modal synthetic data and create positive pairs based on the features of stand-alone RGB and attention-fused RGB. This approach avoids the large discrepancy of using different modalities during pre-training and facilitates better RGB features.

2.2. Semi-Supervised Learning

Due to the difficulty in obtaining real-world 3D ground truth, pose estimation works often study how to learn with limited annotations. Consistency constraints and pseudo-labelling strategies, including shape consistency [9, 13], temporal consistency [5], temporal pseudo-labels [21], and multiview consistency [25, 30] can exploit unlabelled data. Template-corrected pseudo-labels [32] and photometric consistency [8] based on model-fitting can further remove sequence or multi-view requirements.

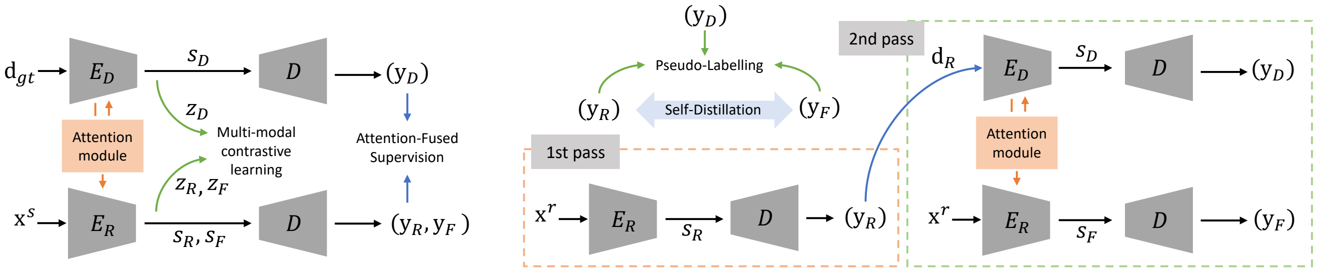
A special case of semi-supervised learning learns from only labelled synthetic data and unlabelled real-world data. This new setting is more challenging due to the additional domain gap. As such, recent works [14, 20] add domain adaptation strategies in addition to training with consistency or pseudo-labels [22, 32]. In this paper, we emphasize using multi-modal data in a semi-supervised setting and leverage the different modalities via feature fusion and alignment. In contrast to [3], which uses ground truth 2D pose and real-world depth maps, we generate other modalities exclusively from real RGB images as pseudo-labels.

3. Architecture

We first introduce our proposed attention module in Sec. 3.1 before presenting the pre-training and fine-tune architectures in Sec. 3.2 that incorporate the attention module for multi-modal learning (see Fig. 2).

3.1. Attention Module

We propose an attention module $\mathrm{Att}(\cdot)$ to estimate local attention weights that capture relationships between local and global feature responses. Suppose $\boldsymbol{f}^R$ and $\boldsymbol{f}^D$ are a pair of corresponding intermediate feature maps from RGB images and depth maps respectively.Given a depth map feature map $\boldsymbol{f}^D \in \mathbb{R}^{c\times h\times w}$, $\boldsymbol{f}^D_{ij} \in \mathbb{R}^{c\times 1\times 1}$ is the feature vector at position $[i, j]$. The average 2D spatial values on $\boldsymbol{f}^D$ are defined as $\bar{\boldsymbol{f}}^D = pool(\boldsymbol{f}^D) \in \mathbb{R}^{c\times 1\times 1}$, where $pool(\cdot)$ is a channel-wise average pooling operation. The attention weight $\boldsymbol{w}$ is defined as the scale-normalized inner product



(a) Pre-training on labeled synthetic multi-modal data

(b) Fine-tuning on unlabelled real-world RGB data

Figure 2. E_R and E_D are encoders for RGB and depth maps; D is a common multi-modal decoder that predicts segmentation masks, depth maps and 2.5D poses. If RGB and depth maps are available (like in pre-training (a)), attention maps derived from the depth encoding in E_D (orange arrows) are applied to the RGB encoder E_R to produce fused features s_F. If there are no depth maps (like in the first pass of fine-tuning (b)), the RGB encoder E_R works stand-alone to produce RGB features s_R. After the first pass in fine-tuning, the decoder D predicts a depth map, which gets applied together with the original RGB input for a second pass to enable self-distillation.

$\langle\cdot\rangle$ of $\bar{\boldsymbol{f}}^D$ and $\boldsymbol{f}_{ij}^D$:

$$\boldsymbol{w}_{ij} = \frac{\langle \bar{\boldsymbol{f}}^D, \boldsymbol{f}_{ij}^D \rangle}{\sum_{i=1}^{h}\sum_{j=1}^{w} \langle \bar{\boldsymbol{f}}^D, \boldsymbol{f}_{ij}^D \rangle}(h \times w). \tag{1}$$

The scaling factor $(h \times w)$ ensures that the attention weights remain within an effective range. The weight $\boldsymbol{w}_{ij}$ is large when the response of position $[i, j]$ is highly correlated with the global response.

Note we do not compute attention weights on RGB image because of inherent ambiguities in the RGB attention weights (see Fig. 1). As such, there are two types of attention-activated outputs: attention-fused RGB features $\text{Att}(\boldsymbol{f}^D) \odot \boldsymbol{f}^R$, or self-attended depth map features $\text{Att}(\boldsymbol{f}^D) \odot \boldsymbol{f}^D$, where $\odot$ is a channel-wise multiplication and $\text{Att}(\boldsymbol{f}^D)$ denotes the weight $\boldsymbol{w}_{ij}$ after inflating along the channel dimension. Additional details are given in the Supplementary.

3.2. Dual-Modality Network

As shown in Fig. 2, the depth map and RGB inputs each have their own encoders, namely E_D and E_R, while sharing a common decoder D. The E_D takes depth maps $\mathbf{d}$ as input and outputs latent features $\boldsymbol{s}_D$. The E_R takes RGB images $\mathbf{x}$ as input and outputs latent features $\boldsymbol{s}_R$. With the attention from depth maps, the E_R also outputs attention-fused features $\boldsymbol{s}_F$. Two fully-connected layers project $\{\boldsymbol{s}_D, \boldsymbol{s}_R, \boldsymbol{s}_F\}$, to 128-dimensional normalized features $\{\boldsymbol{z}_D, \boldsymbol{z}_R, \boldsymbol{z}_F\}$. The decoder D is multimodal and makes a joint predict $\mathbf{y}$ that includes depth maps, segmentation masks, *etc*., from any of the latent features $\boldsymbol{s}_D$, $\boldsymbol{s}_R$ or $\boldsymbol{s}_F$.

The two encoders use a ResNet-101 backbone; the shared decoder features three deconvolution layers with BN and ReLU. Between the two encoders, we embed our proposed attention modules at the end of downsampling layers conv1 and conv2_x to conv5_x. The first set of attention modules are applied for self-attention on the depth encoder, *i.e*. to produce $\text{Att}(\boldsymbol{f}^D) \odot \boldsymbol{f}^D$. The second set of attention modules are applied from the depth encoder to the RGB encoder, to produce attention-fused RGB features $\text{Att}(\boldsymbol{f}^D) \odot \boldsymbol{f}^R$. If depth inputs are not available *e.g*. during testing, or the first pass of the fine-tuning, the original RGB latent features $\boldsymbol{s}_R$ are used to generate outputs.

Training consists of a pre-training and fine-tuning stage. Pre-training is done with paired synthetic RGB and depth map images (Fig. 2 (a)), while fine-tuning is done with real-world RGB images that do not have accompanying depth maps (Fig. 2 (b)). Fine-tuning requires two passes. In the first pass, the input RGB is applied through E_R and D to predict a depth map. In the second pass, the same RGB input is then applied together with the predicted depth map for a second pass. As the attention module activates RGB features in relevant regions, the two passes can be considered a denoising process that refines predictions. During testing, we use RGB images with RGB features only for prediction.

4. Method

4.1. Preliminaries

Suppose we have a set of synthetic data $\mathcal{D}^s = \{(\mathbf{x}_i^s, \mathbf{y}_i)\}_{i=1}^{N_s}$, where each synthesized RGB image $\mathbf{x}_i^s \in \mathbb{R}^{3\times H\times W}$ has a multi-modal label $\mathbf{y}_i = (\mathbf{p}_i, \mathbf{d}_i, \mathbf{m}_i)$ in the form of a 2.5D pose $\mathbf{p}_i \in \mathbb{R}^{J\times 3}$, a depth map $\mathbf{d}_i \in \mathbb{R}^{1\times H\times W}$, and a binary segmentation mask $\mathbf{m}_i \in \mathbb{R}^{1\times H\times W}$. Note that the 2.5D pose $\mathbf{p}$ is expressed as a triplet of the 2D pose and the metric depth relative to the root. Additionally, we have real-world data $\mathcal{D}^r = \{(\mathbf{x}_j^r)\}_{j=1}^{N_r}$, where the real RGB image $\mathbf{x}_j^r \in \mathbb{R}^{3\times H\times W}$ has no labels of any form.

Synthetic-Only Baseline A straightforward approach is to train a model on synthetic data $\mathcal{D}^s$ and then use the model to make predictions on the real-world data $\mathcal{D}^r$. We consider this a *'baseline'*, and use a multi-modal pose estimation pipeline similar to [11] to simultaneously predict

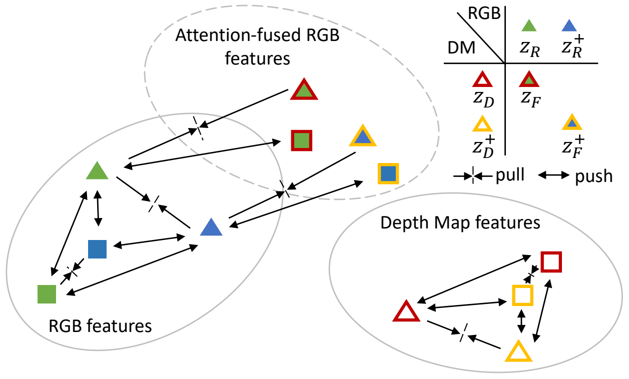


Figure 3. Multi-modal contrastive learning. The triangle and the square represent the normalized features of different poses. In each modality, specific augmentations are used for creating positive pairs, each represented by the same shape but different colors, *e.g.*, triangles in green and blue form a positive pair in RGB. And the augmented sample is denoted with "+". We construct attention-fused RGB features from the raw and augmented RGB-depth map pairs, whose source features can be distinguished from the outline and fill colors. For clarity, we visualize only the complete set of pairwise relationships between the triangular RGB features and attention-fused RGB features. In the top right-hand corner, we provide an example to show the relationship.

the 2.5D representations $\mathbf{p}$, segmentation masks $\mathbf{m}$ and depth maps $\mathbf{d}$ from a given RGB input. For learning, the supervised multi-modal loss is applied with ground truth $\mathbf{y}_{gt} = (\mathbf{p}_{gt}, \mathbf{m}_{gt}, \mathbf{d}_{gt})$ and the corresponding predictions $\mathbf{y}_R = (\mathbf{p}_R, \mathbf{m}_R, \mathbf{d}_R)$:

$$\begin{aligned} M(\mathbf{y}_{gt}, \mathbf{y}_R) = \ell(\mathbf{p}_R, \mathbf{p}_{gt}) + \lambda_{\mathbf{m}} ||\mathbf{m}_R - \mathbf{m}_{gt}||_1 \\ + \lambda_{\mathbf{d}} ||\mathbf{d}_R - \mathbf{d}_{gt}||_1, \end{aligned} \tag{2}$$

where $\lambda_{\mathbf{m}}$ and $\lambda_{\mathbf{d}}$ are weighting hyperparameters. ℓ is the 2.5D pose distance, which is the sum of the weighted Euclidean distance between two 2D poses and that between two metric depths relative to the root keypoint, defined in [32].

Unsurprisingly, models which only train with Eq. 2 do not perform well on real-world data due to the inherent domain gap. As such, we design a dual-modality network with the designed attention module and introduce a pre-training and fine-tuning strategy to overcome the domain gap.

4.2. Pre-training with Multi-level Alignment

Our model is pre-trained with labelled synthetic data $\mathcal{D}^s$ which has ground truth RGB images, masks, depth maps and 2.5D poses. To further leverage the task-specific information in the synthetic depth maps, we align the features from the two encoders with the multi-modal contrastive learning and attention-fused supervision.

Multi-modal Contrastive Learning Based on multi-modal data, we simultaneously perform intra-modal and inter-modal contrastive learning. Intra-modal contrastive learning encourages a discriminative feature space in each modality and is applied to RGB and depth modalities individually. Inter-modal contrastive learning however performs alignment across the RGB and depth modalities.

For intra-modal contrastive learning, we adopt an augmentation strategy to construct contrastive pairs for the RGB modality and the depth map modality individually. Specifically, let $\mathrm{T}_{\mathrm{RGB}}(\cdot)$, $\mathrm{G}_{\mathrm{RGB}}(\cdot)$, $\mathrm{T}_{\mathrm{DM}}(\cdot)$ and $\mathrm{G}_{\mathrm{DM}}(\cdot)$ denote texture and geometric augmentations for RGB and depth respectively. Texture augmentations do not affect the labels, *i.e.*, the hand pose, while geometric augmentations require the labels or hand poses to be adjusted accordingly. These augmentations yield the positive pairs $(\mathbf{x}, \mathrm{T}_{\mathrm{RGB}}(\mathbf{x}))$ or $(\mathrm{G}_{\mathrm{RGB}}(\mathbf{x}), \mathrm{T}_{\mathrm{RGB}}(\mathrm{G}_{\mathrm{RGB}}(\mathbf{x})))$ for the RGB image $\mathbf{x}$ and similarly for the depth maps. The positive pair is passed through the encoder and the projection layers, to yield normalized features $\{\boldsymbol{z}_R, \boldsymbol{z}_R^+\}$ for RGB and $\{\boldsymbol{z}_D, \boldsymbol{z}_D^+\}$ for depth maps. For better understanding, we visualize all push-and-pull operations for intra-domain contrastive learning of the depth map in Fig. 3.

For inter-modal contrastive learning, the naive approach is to form positive samples from each RGB and depth map pair and negative samples from cross-pair combinations. However, there is a large visual difference between RGB images and depth maps. As such, bringing corresponding $\boldsymbol{z}_R$ and $\boldsymbol{z}_D$ close together is not only difficult but ultimately unhelpful for uniformity purposes, *i.e.*, preserving maximal information [31]. Therefore, instead of making each RGB-depth map pair as positive, we use their normalized fused feature $\boldsymbol{z}_F$ for inter-modal contrastive learning with RGB modality. Meanwhile, if the input pair is augmented, the generated normalized fused feature is $\boldsymbol{z}_F^+$. As shown in Fig. 3, we only optimize distances between RGB and attention-fused RGB features (*i.e.*, $\boldsymbol{z}_R$ and $\boldsymbol{z}_F$, $\boldsymbol{z}_R^+$ and $\boldsymbol{z}_F^+$), under the rationale that the attention-fused features are easier to align.

In practice, we adopt the normalized temperature-scaled cross-entropy (NT-Xent) loss as below:

$$\eta(\boldsymbol{z}, \boldsymbol{z}^+) = \\ -\sum_{i=1}^{B} \log \frac{e^{\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_i^+)/\tau}}{\sum_{k=1}^{B} \mathbb{1}_{[k \neq i]} (e^{(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)} + e^{(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k^+)/\tau)})}, \tag{3}$$

where B is the batch size and $\boldsymbol{z}$ and $\boldsymbol{z}^+$ comprise a positive pair of the same sample. The temperature is set to τ=0.5, $\mathrm{sim}(\cdot, \cdot)$ is the cosine similarity, and $\mathbb{1}$ is the indicator function. Based on the NT-Xent loss, we define the contrastive loss for RGB, depth map and fusion as:

$$\begin{aligned} \mathcal{L}_c^{\mathrm{RGB}} = \eta(\boldsymbol{z}_R, \boldsymbol{z}_R^+), \quad \mathcal{L}_c^{\mathrm{DM}} = \eta(\boldsymbol{z}_D, \boldsymbol{z}_D^+) \\ \mathcal{L}_c^{\mathrm{Fusion}} = \eta(\boldsymbol{z}_R, \boldsymbol{z}_F) + \eta(\boldsymbol{z}_R^+, \boldsymbol{z}_F^+). \end{aligned} \tag{4}$$

We formulate the final multi-modal contrastive loss as

$$\mathcal{L}_c = \mathcal{L}_c^{\text{RGB}} + \mathcal{L}_c^{\text{DM}} + \mathcal{L}_c^{\text{Fusion}}. \tag{5}$$

Attention-Fused Supervision As RGB image-depth map pairs have pixel-level correspondences, we propose aligning those features with their attention-fused features that share the same encoder with RGB features. The alignment helps RGB encoder limit its sensitivity to non-informative cues such as background or irrelevant texture. To that end, we apply multi-modal supervision for RGB images, depth maps, and their fusion simultaneously, which propagates the label supervision to the alignment of feature maps. The final attention-fused multi-modal supervised loss is

$$\mathcal{L}_s = M(\mathbf{y}_R, \mathbf{y}_{gt}) + M(\mathbf{y}_D, \mathbf{y}_{gt}) + M(\mathbf{y}_F, \mathbf{y}_{gt}), \tag{6}$$

where $\mathbf{y}_R, \mathbf{y}_D, \mathbf{y}_F$ are the predictions of RGB features, depth map features and attention-fused RGB features, respectively and M is the multi-modal supervised loss defined in Eq. 2.

Overall, we pre-train the model on synthetic data with the following objective function and a hyper-parameter λ_c:

$$\mathcal{L}_{pretrain} = \mathcal{L}_s + \lambda_c \mathcal{L}_c. \tag{7}$$

4.3. Fine-tuning with Noisy Pseudo-Labels

Fine-tuning with real-world data $\mathcal{D}^r$ improves model generalization. As $\mathcal{D}^r$ is unlabelled, we rely on noisy pseudo-labels generated by the pre-trained model. The fine-tuning requires two passes of the data to enable the self-distillation procedure.

4.3.1 Pseudo-Labelling

A standard approach to incorporate real-world data is to predict the unlabelled samples' pseudo-labels, refine the pseudo-labels and add the samples with high confidence into the training set. We follow this approach and use the pseudo-labelling strategy of [32], which features a pose correction to refine the pose labels. The pose correction $\rho(\cdot)$ rectifies bone lengths and joint angles to guarantee biomechanical feasibility of the hand poses.

As shown in Fig. 2 (b), our dual-modality network works in a two-pass manner. We are given $\mathbf{y}_R = (\mathbf{p}_R, \mathbf{m}_R, \mathbf{d}_R)$ from the first pass and $\mathbf{y}_F = (\mathbf{p}_F, \mathbf{m}_F, \mathbf{d}_F)$ from the second pass. The pseudo-labels $\boldsymbol{r}$ are generated by averaging the poses before and after pose correction from $\mathbf{p}_R$ and $\mathbf{p}_F$:

$$\boldsymbol{r} = \frac{1}{4}(\mathbf{p}_R + \rho(\mathbf{p}_R) + \mathbf{p}_F + \rho(\mathbf{p}_F)). \tag{8}$$

Then, we use the pseudo-labels $\boldsymbol{r}$ to supervise all the pose predictions $\mathbf{p}$, as below

$$\mathcal{L}_l = \mathbb{1}(\mathcal{C}(\boldsymbol{r}) \leq \varepsilon)\ell(\mathbf{p}, \boldsymbol{r}), \tag{9}$$

Algorithm 1 Dual-Modality Network.

```
Require: Synthetic data $\mathcal{D}^s$ and real data $\mathcal{D}^r$
Ensure: Final model
for $t = 1, \dots, T_{pretrain}$ epochs do
    Generate augmented samples based on $\mathbf{x}^s$, $\mathbf{d}_{gt}$ from $\mathcal{D}^s$ for $E_R$ and $E_D$
    Calculate $\boldsymbol{s}_R$, and $\boldsymbol{s}_F$ from $E_R$ and $E_D$; Calculate self-attended $\boldsymbol{s}_D$ from $E_D$
    Calculate $\boldsymbol{z}_R$,$\boldsymbol{z}_R^+$,$\boldsymbol{z}_D$,$\boldsymbol{z}_D^+$,$\boldsymbol{z}_F$, $\boldsymbol{z}_F^+$ for contrastive loss via $\boldsymbol{s}_R$, $\boldsymbol{s}_D$, $\boldsymbol{s}_F$ and their projection layers
    Calculate $\mathbf{y}_R$,$\mathbf{y}_D$,$\mathbf{y}_F$ for attention-fused supervision via $\boldsymbol{s}_R$, $\boldsymbol{s}_D$, $\boldsymbol{s}_F$ and $D$
    Update the model via gradient descent of Eq. 7
end for
for $t = 1, \dots, T_{finetune}$ epochs do
    Calculate $\mathbf{y}_R$ with $\mathbf{d}_R$ via $\mathbf{x}^r$ from $\mathcal{D}^r$ and ($E_R$, $D$)
    Calculate $\mathbf{y}_F$ and $\mathbf{y}_D$ via $\mathbf{x}^r$, $\mathbf{d}_R$, ($E_R$, $D$) and ($E_D$, $D$)
    Generate $\boldsymbol{r}$ in Eq. 8 via $\mathbf{y}_R$, $\mathbf{y}_F$ and $\rho(\cdot)$
    Update the model via gradient descent of Eq. 11
end for
```

where $\mathcal{C}(\cdot)$ provides the confidence of the pseudo-labels based on the variance of the poses before and after pose correction in Eq. 8 and ε is a confidence threshold. See Supplementary for more details.

4.3.2 Self-Distillation

Pseudo-labels are inevitably noisy and may change dramatically during fine-tuning, all of which hurt model performance [20]. To address noisy pseudo-labels, our goal is to improve pseudo-labels gradually. As the two-pass dual-modality network can be considered a denoising process, that refines predictions, we exploit RGB images with the predicted depth maps as input to construct a self-distillation structure [17, 20] (See Fig. 2 (b)). By encouraging the refined prediction to be consistent with its original prediction, we distill the knowledge to obtain a softer prediction and generate a pseudo-label accordingly. This strategy helps distill knowledge within the network itself and improves pseudo-labels gradually. Concretely, we apply consistency to the outputs by using multi-modal supervised loss:

$$\mathcal{L}_d = M(\mathbf{y}_R, \mathbf{y}_F), \tag{10}$$

where $\mathbf{y}_R$ are the predictions from RGB and $\mathbf{y}_F$ are the predictions from attention-fused RGB features which take as input RGB images and predicted depth maps $\mathbf{d}_R$.

Overall, we fine-tune the pre-trained model using RGB images of real-world data based on self-distillation $\mathcal{L}_d$ and pseudo-labelling $\mathcal{L}_l$, together with supervision from synthetic data $\mathcal{L}_s$. The overall objective of this stage with hyper-parameters λ_d and λ_l is as follows:

$$\mathcal{L}_{finetune} = \mathcal{L}_s + \lambda_d \mathcal{L}_d + \lambda_l \mathcal{L}_l. \tag{11}$$

The entire pre-training and fine-tuning procedures of our dual-modality network are summarized in Algorithm 1.

5. Experiments

5.1. Datasets & Evaluation

For the model training, we use the synthetic RHD [41], and four real-world hand datasets: STB [38], FreiHAND [42], H3D [39] and MVHand [36]. **RHD** is a large-scale synthetic hand dataset containing 20 characters performing 39 actions. **STB** has 12 video sequences with a total of 15k frames for training and 3k frames for testing. **FreiHAND** is a challenging hand dataset with 130k training images and 4k testing images. In **H3D**, we consider the subset of one-handed gestures for our experiments, which comprises 11k training data and 2k testing data. **MVHand** is a newly released hand dataset which has 42k images for training and 42k images for testing.

We evaluate 2D keypoint detection with the percentage of correct keypoints (PCK) and regard an estimation as correct when its distance to the ground truth is within 0.05 of the output size. For 3D pose estimation, we evaluate accuracy via the mean end-point-error (EPE) and the area under the curve (AUC) of the 3D PCK.

5.2. Implementation Details

We adopt ResNet-101 initialized from ImageNet as the backbone. The input data is cropped around the hand and resized to 256 × 256, while the resolution for the output is 64 × 64. We pre-train on synthetic data using the Adam optimizer with a momentum of (0.9, 0.99) and a learning rate of 2.5e-4 that gets decreased by a factor of 0.1 after 40 epochs and 50 epochs. We then fine-tune on synthetic and real data with a learning rate of 2.5e-5 for 6 epochs. Batch size is 140 for pre-training and 20 for fine-tuning. The hyper-parameters of Eqs. 2 and 7-11 are set empirically with $\lambda_\mathbf{m} = 100$, $\lambda_\mathbf{d} = 50$, $\lambda_c = 0.1$, $\lambda_f = 0.2$, $\lambda_l = 1$ and $\varepsilon = 1.5$.

For synthetic data $\mathcal{D}^s$ and pre-training, we use RHD. The other real-world datasets are considered for $\mathcal{D}^r$; as per [32], we fine-tune with a single real-world dataset's training data and report results on the evaluation data. On FreiHAND and MVHand, we select only a subset of the training to match the size of the STB training set. The results of fine-tuning using the full training set are described in the Supplementary Material regarding the impact of real-data volume. Empirical results of the compared methods are directly reported if available; otherwise, they are generated based on officially released code.

PCK@0.05	STB	FreiHAND	H3D	MVHand
Baseline	0.547	0.511	0.555	0.515
RegDA [14]	0.613	0.622	0.720	0.601
CC-SSL [22]	0.655	0.631	0.717	0.602
AnimalDA [20]	0.631	0.629	0.676	0.640
SemiHand [32]	0.668	0.564	0.672	0.563
Ours	**0.775**	**0.658**	**0.749**	**0.689**
Improvement	↑16.0%	↑4.3%	↑4.0%	↑7.7%

Table 1. The performance comparisons for 2D keypoint detection. The relative performance boost between 1st and 2nd best methods can be seen in *Improvement*. Our approach brings consistent improvements over previous state-of-the-art methods. **Bold** numbers indicate the best performance.

EPE(mm)	STB	FreiHAND	H3D	MVHand
Baseline	19.66	21.56	27.77	21.21
SemiHand [32]	14.60	19.33	19.19	19.75
+DM weak labels	13.74	18.71	19.02	18.95
+DM encoder	**13.17**	**18.32**	**18.65**	**18.53**
Ours	**11.99**	**15.61**	**17.08**	**16.45**
Improvement	↑9.0%	↑14.8%	↑8.4%	↑11.2%

Table 2. The performance comparisons with SemiHand and its variants for 3D keypoint estimation. The relative performance boost between 1st and 2nd best methods can be seen in *Improvement*. Adding a depth map (DM) yields improvements for SemiHand, but cannot surpass our method. **Bold** numbers indicate the best performance.

5.3. 2D Keypoint Detection

Table 1, compares our approach with state-of-the-art methods [14,20,22,32] for 2D keypoint detection tasks. All compared methods outperform the multi-modal supervision baseline (Eq. 2) trained with only RHD supervision, though ours is the best for all four benchmarks. Compared to SemiHand [32], which is the most related work, we improve the performance for FreiHAND and MVHand by **16.6%** (SemiHand's 0.564 vs. our 0.658) and **22.4%** (SemiHand's 0.563 vs. our 0.689), respectively. Our results verify that 2D keypoint detection benefits from learning with depth maps to focus on areas with rich geometric information and semantic meaning.

5.4. 3D Keypoint Estimation

SemiHand is the only comparable published work. For fairness of comparison, we also add depth maps to SemiHand baseline, either as weak supervision on the output or as depth map encoder. Table 2 shows that both SemiHand and our approach outperform the baseline, though our method further decreases the best variant of Semihand's EPE up to 2.7mm (18.32mm vs. 15.61mm). The decrease is larger for the datasets with complex backgrounds and more camera views (*i.e.*, FreiHAND and MVhand) confirming our aim of forcing the model to focus on semantically meaningful areas.

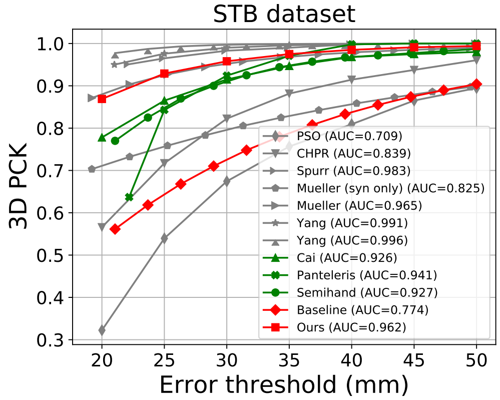


Figure 4. Comparisons with state-of-the-art on STB. Our method achieves the highest AUC score among weakly/semi-supervised methods (green lines), while having comparable performance to other supervised learning methods.

	Method	STB	FreiHAND	H3D	MVHand
Res101	Baseline	19.66	21.56	27.77	21.21
	ℓ_1 alignment [37]	16.38	21.58	26.00	21.58
	Intra-CL [6]	17.32	19.95	26.33	20.40
	Intra-/inter-CL [28]	18.01	19.49	26.61	19.71
	Ours	**16.37**	**19.19**	**25.94**	**19.62**
Res50	Baseline	20.47	23.57	34.42	23.84
	PeCLR [26]	18.84	22.17	31.69	21.96
	Ours	**17.73**	**19.94**	**26.91**	**20.65**

Table 3. Comparison for 3D keypoint estimation for pre-training strategies for RHD. Our model leverages the depth map modality and achieves the best cross-domain performance. **Bold** numbers indicate the best performance.

Fig. 4 shows the AUC curves on 3D PCK to compare with other state-of-the-art. Our proposed method surpasses all weakly- and semi-supervised methods [2, 24, 32] by a large margin on STB dataset. Our performance is comparable to other supervised learning methods [23,27,33,35,41].

5.5. Analysis on Pre-training

Table 3 compares our pre-training strategy with state-of-the-art methods. We first apply ℓ_1 alignment between feature maps of RGB and depth maps [37]. The results show that it fails to handle FreiHAND and MVHand. We further explore contrastive learning to align features across modalities in low-dimensional latent spaces. There is a slight improvement when using intra-CL [6, 40], *i.e.*, $\mathcal{L}_c^{\text{RGB}} + \mathcal{L}_c^{\text{DM}}$ in each modality. Additionally, we follow [28] and apply intra-CL with direct inter-CL (denoted as intra-/inter-CL), *i.e.* $\mathcal{L}_c^{\text{RGB}}+\mathcal{L}_c^{\text{DM}}+\eta(\boldsymbol{z}_R, \boldsymbol{z}_D)$, but this causes accuracy drops on STB and H3D compared to intra-CL. It is likely that the large visual differences between the RGB and depth map result in negative latent feature pairs which are already distant and lead to invalid uniformity effects [29].

In contrast, our multi-modal contrastive learning is not compromised by the large discrepancy between the RGB image and depth map since we contrast with the attention-fused RGB features. Compared to depth map features, attention-fused RGB features are considerably closer to RGB features. This is because attention-fused RGB features are only refined RGB features with the geometric information of depth maps, and the network can learn this information by aligning the two features. As such, our constructed feature pairs with specific properties are more suitable for contrastive learning.

Method	STB	FreiHAND	H3D	MVHand
$M(\mathbf{y}_R, \mathbf{y}_{gt})$	19.66	21.56	27.77	21.21
$+M(\mathbf{y}_D, \mathbf{y}_{gt})$	17.83	20.14	26.56	20.30
$+M(\mathbf{y}_F, \mathbf{y}_{gt})$	16.92	19.69	26.27	19.91
$+\mathcal{L}_c$	16.37	19.19	25.94	19.62
$+\mathcal{L}_l$	12.59	16.27	17.45	16.91
$+\mathcal{L}_d$	**11.99**	**15.61**	**17.08**	**16.45**

Table 4. Ablation study for the input modalities and the components of our approach. Adding (+) the modalities and the components incrementally improves performance. **Bold** numbers indicate the best performance.

The experimental results on model pre-training also verify the effectiveness of our method which consistently outperforms the state-of-the-art methods [28, 40] on all the benchmarks.

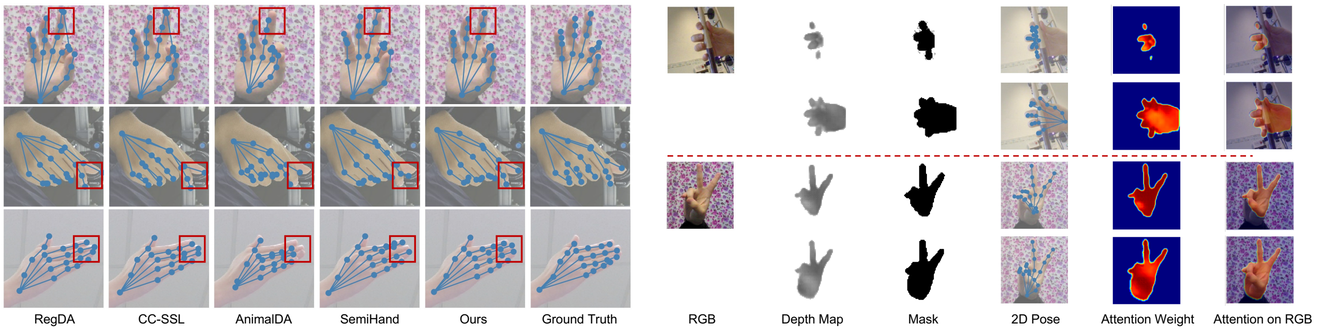
We also compare PeCLR [26] that targets equivariant contrastive learning with geometric augmentations for pose estimation. For a fair comparison, we use a ResNet-50 backbone as per PeCLR, and pre-train both PeCLR and our network on RHD according to our supervised contrastive learning setting. Our method consistently outperforms PeCLR in the four datasets, confirming the effectiveness of our multi-level alignment in pre-training.

5.6. Ablation Study

In Table 4, we ablate the impact of input modalities and our proposed components. Due to the architectural dependencies, we can only perform the ablation study in an incremental manner. The importance of adding depth maps for learning is verified by the significant accuracy boost when comparing "$M(\mathbf{y}_R, \mathbf{y}_{gt})$" with "$+M(\mathbf{y}_D, \mathbf{y}_{gt})$". Applying the multi-modal supervised loss on their fusion ($+M(\mathbf{y}_F, \mathbf{y}_{gt})$) further improves the accuracy. The contributions of multi-modal contrastive learning ($+\mathcal{L}_c$), pseudo-labelling ($+\mathcal{L}_l$), and self-distillation ($+\mathcal{L}_d$) are outlined in Table 4. The incremental addition of the components successively decreases the mean EPE for all four datasets.

5.7. Qualitative Results

Fig. 5 (a) compares examples of the 2D keypoint detection on STB, H3D and MVHand. Our predictions are most similar to the ground truth, while the other methods have



(a) Comparison of 2D keypoint detection

(b) Visualization before and after fine-tuning

Figure 5. (a) 2D keypoint visualization on STB, H3D and MVHand. We compare our method with four state-of-the-art methods and highlight the differences between the predictions and the ground truth poses with red boxes. (b) visualization of two examples before fine-tuning (first row) and after fine-tuning (second row). From left to right: RGB images, multi-modal predictions (depth maps, segmentation masks, poses), attention weights from depth maps and attention weights on RGB images. Figure best viewed in color.

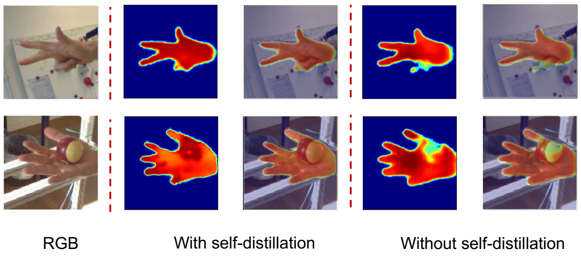


Figure 6. Comparisons of attention with and without self-distillation. The proposed self-distillation can optimize the generated attention map to encourage better activation on the hand.

wrong predictions, especially at the fingertips.

Fig. 5 (b) visualizes the predicted depth maps and the attention weights from conv1 layers before and after model fine-tuning. The FreiHAND sample (top) features object occlusion, while the STB sample (bottom) has extreme lighting. In both cases, the pre-trained model can only estimate the depth values for a few areas, *e.g.*, the fingers, or misses the palm. But after fine-tuning, the model successfully estimates depth maps with complete hands. The attention weights before and after directly reveal the activation of missing areas after fine-tuning.

In Fig. 6, we also show the comparison of attention maps with and without self-distillation. If we exclude self-distillation in our method, the attention map cannot promise high activation on full hand. After adding self-distillation, we can alleviate the problem of low activation caused by occlusion. See the Supplementary Material for more qualitative results.

5.8. Failure Cases

The first row of Fig. 7 shows some failure cases where the multi-modal outputs are inconsistent, *e.g.*, despite complete depth maps, the 2D pose is still incorrect. The second row of Fig. 7 shows failure cases when estimating occluded keypoints; this challenging setting occurs commonly for hand gestures with self-occlusion and is an ill-posed problem from a single view. As such, multi-modal and multi-view consistencies are possible extensions for future work.

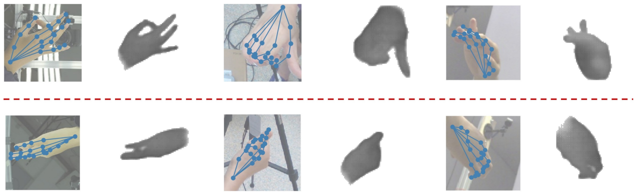


Figure 7. Failure cases with inconsistent cross-modal predictions (first row) and self-occlusion (second row). The samples are taken from H3D (left), MVHand (middle) and FreiHAND (right). Figure best viewed in color.

6. Conclusion

In this paper, we propose a dual-modality network to address the cross-domain pose estimation problem in a semi-supervised setting. By leveraging the multiple data modalities of synthetic data, we explore multi-level alignment during pre-training, including multi-modal contrastive learning and attention-fused supervision. During fine-tuning, we explore self-distillation based on our proposed dual-modality network and provide a unified fine-tuning scheme for real data with noisy pseudo-labels. Our experiments show that our approach significantly outperforms state-of-the-art methods on four datasets. In the future, we intend to dive deeper into contrastive learning and self-distillation for semi-supervised hand pose estimation and explore multi-modal multi-view consistency for hand sequences.

Acknowledgments This research / project is supported by the Ministry of Education, Singapore, under its MOE Academic Research Fund Tier 2 (STEM RIE2025 MOE-T2EP20220-0015).

References

[1] Giuseppe Caggianese, Nicola Capece, Ugo Erra, Luigi Gallo, and Michele Rinaldi. Freehand-steering locomotion techniques for immersive virtual environments: A comparative evaluation. *IJHCI*, 36(18):1734–1755, 2020. 1

[2] Yujun Cai, Liuhao Ge, Jianfei Cai, Nadia Magnenat Thalmann, and Junsong Yuan. 3d hand pose estimation using synthetic data and weakly labeled rgb images. *TPAMI*, 43(11):3739–3753, 2020. 7

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

[4] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *ICCV*, 2019. 2

[5] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, and Xiaohui Xie. Temporal-aware self-supervised learning for 3d hand pose and mesh estimation in videos. In *WACV*, 2021. 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 7

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[8] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *CVPR*, 2021. 2

[9] Zheng Chen, Sihan Wang, Yi Sun, and Xiaohong Ma. Self-supervised transfer learning for hand mesh recovery from binocular images. In *ICCV*, 2021. 2

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2

[11] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression. In *ECCV*, 2018. 3

[12] Seogkyu Jeon, Kibeom Hong, Pilhyeon Lee, Jewook Lee, and Hyeran Byun. Feature stylization and domain-aware contrastive learning for domain generalization. In *ACMMM*, 2021. 2

[13] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, 2021. 2

[14] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *CVPR*, 2021. 1, 2, 6

[15] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *ICLR*, 2020. 2

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 2

[17] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation with progressive refinement of targets. In *ICCV*, 2021. 2, 5

[18] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *TPAMI*, 42(9):2306–2320, 2019. 1

[19] Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. A deep learning approach for analyzing video and skeletal features in sign language recognition. In *IST*, 2018. 1

[20] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *CVPR*, 2021. 2, 5, 6

[21] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, 2021. 2

[22] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *CVPR*, 2020. 2, 6

[23] 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*, 2018. 1, 7

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

[25] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2

[26] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. In *ICCV*, 2021. 2, 7

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

[28] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 2, 7

[29] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *NeurIPS*, 2020. 7

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

[31] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020. 4

[32] Linlin Yang, Shicheng Chen, and Angela Yao. Semihand: Semi-supervised hand pose estimation with consistency. In *ICCV*, 2021. 1, 2, 4, 5, 6, 7

[33] Lixin Yang, Jiasen Li, Wenqiang Xu, Yiqun Diao, and Cewu Lu. Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. In *BMVC*, 2020. 7

[34] Linlin Yang, Shile Li, Dongheui Lee, and Angela Yao. Aligning latent spaces for 3d hand pose estimation. In *ICCV*, 2019. 2

[35] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *CVPR*, 2019. 7

[36] Ziwei Yu, Linlin Yang, Shicheng Chen, and Angela Yao. Local and global point cloud reconstruction for 3d hand pose estimation. In *BMVC*, 2021. 6

[37] Shanxin Yuan, Bjorn Stenger, and Tae-Kyun Kim. Rgb-based 3d hand pose estimation via privileged learning with depth images. *arXiv preprint arXiv:1811.07376*, 2018. 1, 7

[38] 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

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

[40] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In *GCPR*, 2021. 2, 7

[41] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, 2017. 1, 6, 7

[42] 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*, 2019. 6