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

In the following supplementary material, we present more details about our proposed method. Sec. A provides an illustration of the attention structure and the interaction between attention and RGB and depth map encoders. We also compare the performances using different attention sources (*i.e.*, RGB and depth map). Sec. B provides a comprehensive description of the augmentations used in contrastive learning. Sec. C provides an illustration of how we performed pose correction in our work. Sec. D discusses the stop gradient operation in our attention module for attention-fused features. Sec. E shows the influence of data amount on ours and other state-of-the-art methods. Sec. F provides more visualizations about multi-modal predictions, and 2D pose predictions.

A. Attention Module

Architecture Suppose $\boldsymbol{f}^R$ and $\boldsymbol{f}^D$ are a pair of corresponding intermediate feature maps from RGB images and depth maps respectively. In Fig. a, we provide the detailed structure of the proposed attention module. This module is only applied at the end of downsampling layers of E_D and E_R to produce attention-activated features, as shown in Fig. b. The attention is calculated from the depth feature map $\boldsymbol{f}^D$ and applied to itself and the RGB feature map $\boldsymbol{f}^R$, respectively. With the attention guidance, E_R outputs attention-fused latent feature $\boldsymbol{s}_F$ and its original RGB features $\boldsymbol{s}_R$. E_D outputs self-attended features $\boldsymbol{s}_D$.

RGB Attention vs. Depth Map Attention In Table. a, we provide the experiment results of using RGB self-attention and depth map attention to guide the learning of RGB modality. Both methods use the proposed pre-training loss with the same encoder and decoder architectures. The results show that attention from depth map can help RGB better learn task-relevant information.

B. Augmentation

In this section, we introduce the details of different augmentation strategies for contrastive learning.

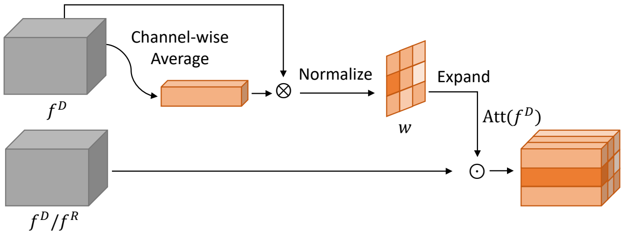


Figure a. Attention module. Attention weights are generated from $\boldsymbol{f}^D$ and are applied to either $\boldsymbol{f}^D$ to produce self-attended depth map features or $\boldsymbol{f}^R$ to produce attention-fused RGB features.

Contrastive Learning For contrastive learning, we define $\mathrm{T_{RGB}}(\cdot)$, $\mathrm{G_{RGB}}(\cdot)$, $\mathrm{T_{DM}}(\cdot)$ and $\mathrm{G_{DM}}(\cdot)$ as the texture and geometric augmentations of the RGB image and depth map, respectively. Texture augmentations do not affect the labels, *i.e.*, the hand poses, while geometric augmentations require the labels or hand poses to be adjusted accordingly. We list the details below:

1. $\mathrm{T_{RGB}}(\cdot)$ consists of colour jitter, grey-scale and random erasure.
2. $\mathrm{G_{RGB}}(\cdot)$ consists of a rotation of [-180°,180°], scale of [0.8,1] and translation of [-20,20] pixels.
3. $\mathrm{T_{DM}}(\cdot)$ consists of random erasure, salt and pepper noise.
4. $\mathrm{G_{DM}}(\cdot)$ consists of a rotation of [-180°,180°], scale of [0.8,1] and translation of [-20,20] pixels.

C. Pose Correction

Here, we introduce the pose correction [5] with a 2D example. As shown in Fig. c(a), we have a given template (solid line) and a prediction (dotted line). For the template, we have four joints, *i.e.*, r, $\{j_i\}_{i=1:3}$ from root to leaf, and three bones $\{b_i\}_{i=1:3}$. Correspondingly, we have the predicted joints $\hat{r}$, $\{\hat{j}_i\}_{i=1:3}$ and predicted bones $\{\hat{b}_i\}_{i=1:3}$. Note that we also define the valid interval for b_1 and b_2,

[*] Equal contribution

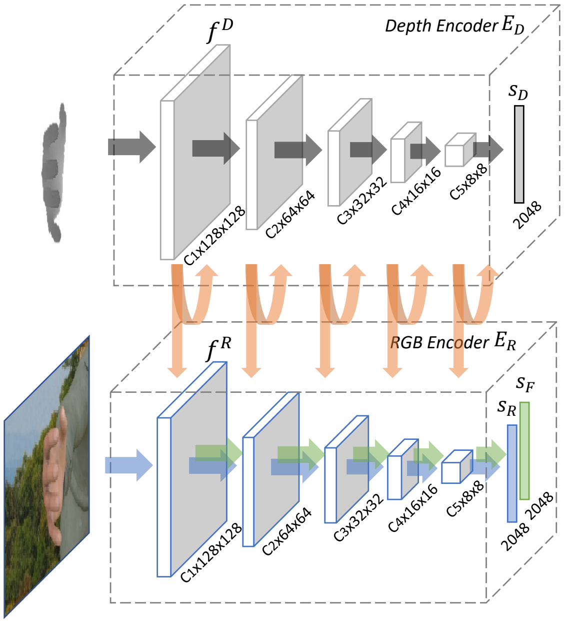


Figure b. Attention w/ encoders. We illustrate the intermediate feature maps of RGB and depth map where attention modules are applied (*i.e.*, downsampling layers with $C_{1,2,3,4,5} = \{64, 256, 512, 1024, 2048\}$). Note that orange streams are the attention deployment; black streams are depth map feature processing; blue streams are RGB feature processing while green streams are the attention-fused RGB feature processing.

as shown in the yellow triangle area. We build a local coordinate system for the hand and use the valid interval of each bone based on the work [4]. Our goal is to register the template to the prediction and get registered joints $\bar{r}$, $\{\bar{j}_i\}_{i=1:3}$. In this case, we use a greedy approximation to avoid the accumulation of endpoint errors and to ensure the feasibility of the hand pose. As shown in Fig. c(b)-(f): our pose correction method consists of the following steps:

1. We first align the root to get $\bar{r}$ by translating r to $\hat{r}$ (Fig. c(b)).

2. We calculate the transformation of b_1 based on the direction from $\bar{r}$ to $\hat{j}_1$, *i.e.*, the grey dotted line in Fig. c(c).

3. We get registered $\bar{j}_1$ based on the transformation of b_1. We also get the transformation of b_2 based on the direction from registered $\bar{j}_1$ to $\hat{j}_2$, as shown in Fig. c(d).

4. We get registered $\bar{j}_2$ based the transformation of b_2. As the joint angle is valid, no rectification is needed. We follow by calculating the transformation of b_3 based on the direction from $\bar{j}_2$ to $\hat{j}_3$, as shown in Fig. c(e).

5. In Fig. c(f), we get registered $\bar{j}_3$ based on the transformation of b_3. Note that the joint angles that exceed a valid interval are rectified. Therefore, we get complete registered poses.

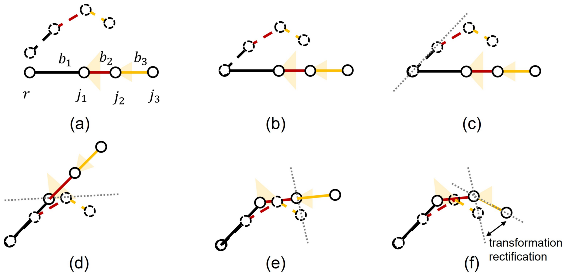


Figure c. Illustration of pose correction. We use a greedy approximation to correct the predicted hand pose based on a hand template to realize bone length and joint angle constraints.

Method	STB	FreiHAND	H3D	MVHand
from RGB	16.92	19.51	26.55	20.27
from Depth Map	**16.37**	**19.19**	**25.94**	**19.62**

Table a. Ablation study for the attention source. We investigate the attention source for the RGB feature by applying RGB self-attention or depth map attention to conduct the proposed model pre-training. The results show that attention maps from depth maps provide better guidance for RGB than attention maps from RGB. **Bold** indicates the best performance.

D. Stop Gradients

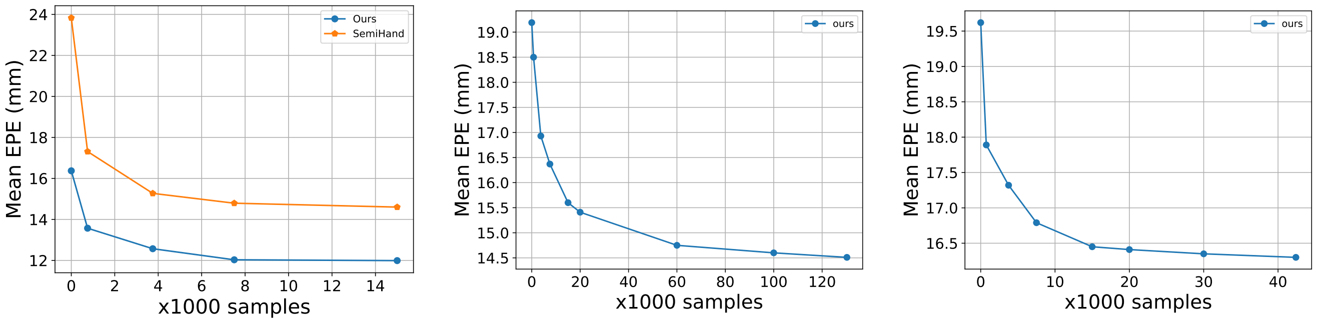
During pre-training, for the attention-fused features, a stop-gradient operation is added between RGB features and attention weights to prevent inaccurate RGB features from degenerating the attention. This is because the attention-fused features alone are insufficient to fully discard distractor information that may be present in the features.

As shown in Table b, we compare our pre-trained models with and without the stop-gradient operation (“w/ SG” versus “w/out SG”). The results “w/ SG” outperform the results “w/out SG” for all four datasets. This confirms our hypothesis that the attention from the depth map encoder makes it easier for the RGB encoder to capture the geometric information and improve cross-dataset performance. On the other hand, the abundant non-informative features captured in the RGB encoder impede the learning of the attention module.

E. Impact of the Amount of Data

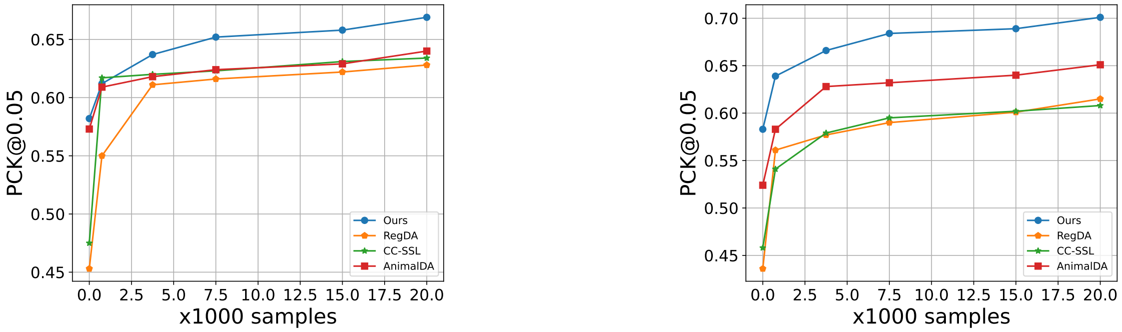
E.1. 3D Keypoint Estimation

Fig. d shows the performance of 3D keypoint estimation using a subset of the training data, from 750 samples to full samples in each dataset. The mean EPE of models



(a) STB (b) FreiHAND (c) MVHand

Figure d. The performance comparisons for 3D keypoint estimation with fine-tuning on different amount of training data. (a) Mean EPE on STB testing data with fine-tuning on different amount of STB training data; Our method can achieve impressive improvement at a higher baseline. (b) Mean EPE on FreiHAND testing data with fine-tuning on different amounts of FreiHAND training data; (c) Mean EPE on MVHand testing data with fine-tuning on different amounts of MVHand training data. The performance saturation can be seen at 60K samples on FreiHAND dataset and 20K samples on MVHand dataset. Figure best viewed in colour.



(a) FreiHAND (b) MVHand

Figure e. The performance comparisons for 2D keypoint detection with fine-tuning on different amounts of training data. (a) PCK@0.05 on FreiHAND testing data with fine-tuning on different amounts of FreiHAND training data (b) PCK@0.05 on MVHand testing data with fine-tuning on different amounts of MVHand training data. Figure best viewed in color.

Method	STB	FreiHAND	H3D	MVHand
w/out SG	17.22	19.55	26.01	20.17
w/ SG	**16.37**	**19.19**	**25.94**	**19.62**

Table b. Ablation study for the stop gradient operation during pre-training. w/ and w/out SG: denotes training with and without the stop-gradient operation between RGB features and attention weights respectively. Our full model ("w/ SG") correctly leverages the depth map modality and generates better representations for the cross-domain dataset. **Bold** indicates the best performance.

on the test set generally improves as the amount of training samples increases. We compare our method with SemiHand on STB as they released their results. As shown in Fig. d(a), with 750 STB training data, SemiHand shows a mean EPE decrease from 23.83 mm to 17.31 mm (**27.4%** improvement). In contrast, our method benefits from the proposed pre-training scheme as we can lower the mean EPE from 19.66 mm to 16.37 mm without using any STB data. We then further reduce the mean EPE to 13.57 mm with 750 STB training samples. In total, we can achieve a **31.0%** improvement over 19.66 mm, which is superior to the **27.4%** improvement given by SemiHand. If we use 15K STB training samples, our method can improve the performance by **39.0%** (19.66 mm to 11.99 mm), while the performance increase of SemiHand is **38.7%** (23.83 mm to 14.60 mm). Although our method is less effective than SemiHand from the aspect of fine-tuning, the improvement is still impressive (from 16.37 mm to 11.99 mm), considering we are at a higher baseline.

We further show the influence of the amount of training data on FreiHAND and MVHand in Fig. d(b)-(c). Our method clearly achieves better performance with more data. Furthermore, we can see that the performance of our method starts to saturate at 60K samples on FreiHAND dataset and 20K samples on MVHand dataset. This phenomenon may be due to the larger pose space in FreiHAND, while MVHand is constructed from streaming data with

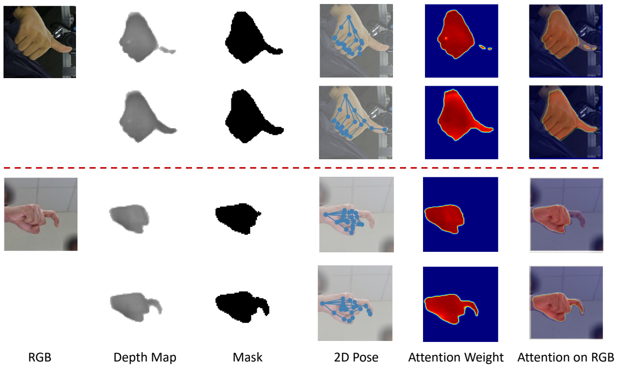


Figure f. Visualization of two examples before fine-tuning (first row) and after fine-tuning (second row). From left to right: RGB images, multi-modal predictions (depth maps, segmentation masks, poses), attention weights from depth maps, and attention weights on RGB images. We can obtain better multi-modal predictions after fine-tuning. Figure best viewed in color.

common poses.

E.2. 2D Keypoint Detection

Fig. e shows the performance comparisons for 2D keypoint detection with fine-tuning on different amounts of training data. We compare our method with the state-of-the-art methods, *i.e.*, RegDA [1], CC-SSL [3] and AnimalDA [2].

As shown in Fig. e (a), for FreiHAND, our method and AnimalDA are both at a similar baseline. However, when using 7.5K samples, we achieve a **12%** improvement, which is superior to the **8.9%** improvement given by AnimalDA. Compared to RegDA and CC-SSL with a lower baseline, even they have an impressive improvement when starting to fine-tune with real-world data. However, their PCK@0.05 gaps to our proposed method are constantly larger than 1.7% when using more than 2.5K samples. Moreover, the performance of our method starts to saturate at 15K samples while others start to saturate at 5K samples on FreiHAND. As for MVHand in Fig. e (b), similar results can be found, where our method outperforms other methods by a large margin and all methods exhibit saturation when using 15k samples in fine-tuning.

Based on these observations, we can conclude that 15K samples are sufficient for fine-tuning to achieve a stable result. For convenience, we choose to use 15K samples, the same size as the STB training set from the FreiHAND and MVHand training set, during fine-tuning.

F. More Qualitative Results

F.1. Multi-Modal Predictions

In Fig. f, we provide the multi-modal prediction changes before and after fine-tuning two examples from H3D and MVHand. It is worth noting that our method can improve the performance, exhibiting more complete depth maps and masks, as well as more accurate 2D poses. This verifies the effectiveness of our proposed fine-tuning strategy.

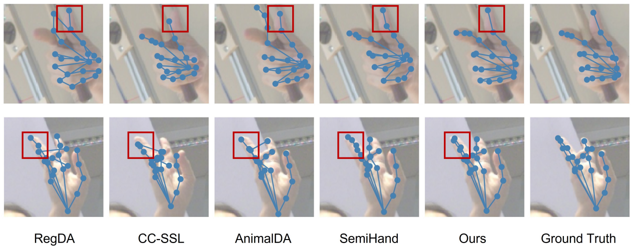


Figure g. 2D pose visualization. We compare our method with four state-of-the-art methods and highlight the differences between the predictions and the ground truth poses with red boxes. Our method can generate more correct predictions. Figure best viewed in color.

F.2. 2D Pose Comparisons

In Fig. g, we provide more 2D pose visualizations. The results show that our method is superior to other state-of-the-art methods and exhibits good performance on complex gestures, which are challenging for other methods.

References

[1] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *CVPR*, 2021. 4

[2] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *CVPR*, 2021. 4

[3] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *CVPR*, 2020. 4

[4] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *ECCV*, 2020. 2

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