

# Supplementary Materials: Learning Context with Priors for 3D Interacting Hand-Object Pose Estimation

Anonymous Authors

This supplementary material includes three sections. Section 1 illustrates the structural details of Customized Feature Maps (CFM) and its variants. In Section 2, we provide a more detailed ablation study of the LCP. Section 3 compares the model complexity between our approach and state-of-the-art methods.

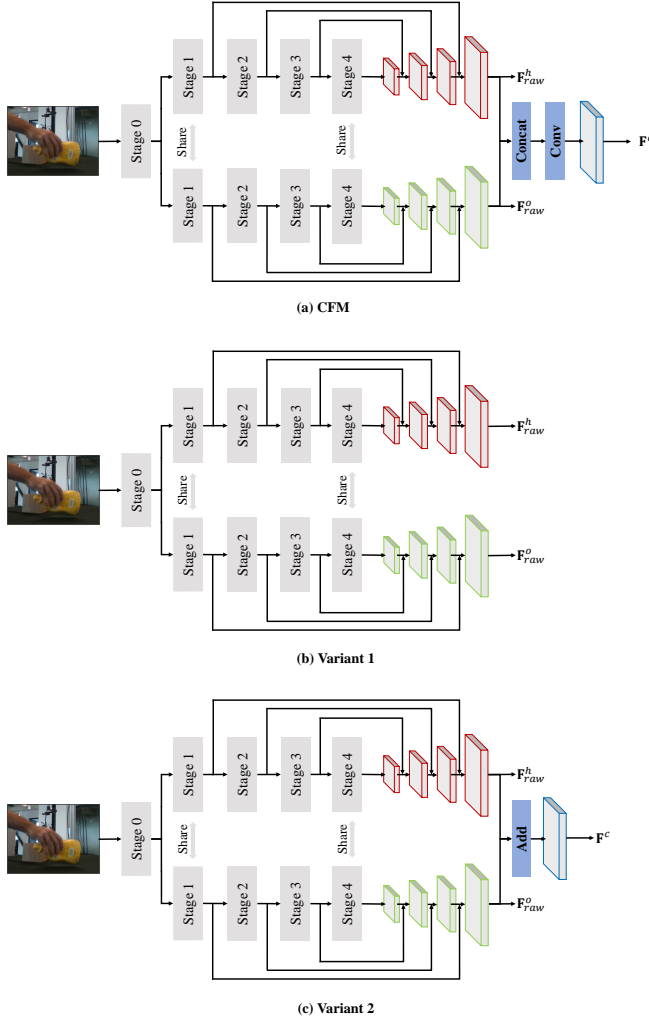


Figure 1: Illustration of the CFM backbone and its variants.

## 1 STRUCTURAL DETAILS OF CFM AND ITS VARIANTS

Fig. 1 illustrates the structure of the CFM backbone and its two variants that were listed in Table 4 of the main paper. CFM is based on the backbone proposed in [2], which not only disentangles the hand and object feature maps but also ensures that they share the

same feature space. The two feature maps are denoted as  $F_{raw}^h \in \mathbb{R}^{H/4 \times W/4 \times C}$  and  $F_{raw}^o \in \mathbb{R}^{H/4 \times W/4 \times C}$ , respectively. To integrate this backbone with LCP, we concatenate  $F_{raw}^h$  and  $F_{raw}^o$  along the channel dimension and then halve the channel number by an efficient  $1 \times 1$  convolution layer. The obtained feature map  $F^c \in \mathbb{R}^{H/4 \times W/4 \times C}$  is utilized as the value and key for the context decoder layer.  $F^h$  and  $F^o$  are obtained by applying ROIAlign [1] to  $F_{raw}^h$  and  $F_{raw}^o$ , respectively.

In comparison, the first variant (Fig. 1 (b)) directly utilizes  $F_{raw}^h$  and  $F_{raw}^o$  in the context decoder layer. In other words, the hand queries utilize  $F_{raw}^h$  as the key and value; while the object queries adopt  $F_{raw}^o$  as the key and value in the context decoder layer.

The second variant (Fig. 1 (c)) replaces the concatenation operation in Fig. 1 (a) with the simple element-wise addition operation to fuse  $F_{raw}^h$  and  $F_{raw}^o$ .

## 2 MORE ABLATION STUDY ON LCP

In Table 1, we compare the performance of LCP with another three variants. The first variant contains four hand and four object decoder layers. It does not include any context decoder layers. The hand and object decoder layers adopt  $F^h$  and  $F^o$  as the value and key in the cross-attention operations, respectively. Therefore, this variant utilizes features in the hand and object bounding boxes only.

Compared with the first variant, the second variant enlarges the cross-attention area for all four hand decoder layers. This means that these hand decoder layers utilize  $F$  as the value and key in the cross-attention operations.

Compared with our LCP model, the third variant adopts more parameter sharing. It further shares the parameters between each hand decoder layer and its counterpart in the object decoder layers.

Table 1: More ablation study on LCP.

Model	Hand		Object
	PA-MPJPE ↓	MPJPE ↓	ADD(-S) ↑ (Average)
LCP	5.33	12.50	49.6
variant 1	5.51	13.37	47.6
variant 2	5.65	13.52	47.9
variant 3	5.40	12.76	46.9

As shown in Table 1, the performance of all the three variants is lower than ours. The comparison between LCP and the first variant indicates that utilizing broader range of context is vital for robust hand-object pose estimation. The comparisons between the first and second variants suggest that it is harmful to extract context features in each hand decoder layer, since context also introduces interference. Additionally, the comparison between the third variant and LCP indicates that sharing parameters between each set of hand and object decoder layers is also harmful, as it hinders the learning of unique hand or object features.

### 3 COMPARISONS IN MODEL COMPLEXITY

We compare the model complexity between state-of-the-art methods and ours on the Dex-YCB database. The average inference speed is calculated on the Dex-YCB test set with an NVIDIA GeForce RTX 3090 GPU using corresponding official codes.

As shown in Table 2, LCP<sup>†</sup> achieves the best hand and object pose estimation accuracy while maintaining a model size comparable to [2] and exhibiting higher frames per second (FPS). These comparisons show that our method is both powerful and efficient.

**Table 2: Comparisons in model complexity and hand-object pose estimation accuracy.**

Method	Params ↓	FPS ↑	PA-MPJPE ↓	MPJPE ↓	ADD(-S) ↑ ( <i>Average</i> )
HFLNet [2]	46.08M	38	5.47	12.56	30.2
LCP <sup>†</sup>	45.36M	43	<b>5.14</b>	<b>11.81</b>	<b>50.6</b>

### REFERENCES

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *ICCV*.  
[2] Zhifeng Lin, Changxing Ding, Huan Yao, Zengsheng Kuang, and Shaoli Huang. 2023. Harmonious Feature Learning for Interactive Hand-Object Pose Estimation. In *CVPR*.