# Supplementary Material for "LGNet: Local-and-Global Feature Adaptive Network for 3D Interacting Hand Mesh Reconstruction"

Haowei Xue[1,2,3] Meili Wang[1,2,3,*]

In this supplementary material, we provide additional experiments, discussions, and other details that could not be included in the main text due to lack of space. The content is summarized as follows:

- Detailed explanation of feature definitions
- Detailed architecture of LGNet
- Detailed experiments
- Discussion

Note that all the notation and abbreviations here are consistent with the main manuscript.

## I. Detailed explanation of feature definitions

We will explain these definitions in detail:

(1) **Local features** (see main manuscript Fig.3: $\mathbf{F_R}$, $\mathbf{F_L}$): refers to intra-hand features, which mainly capture fine-grained details and movements of each hand.

(2) **Interaction features** (see main manuscript Fig. 3: $\mathbf{F}_{\mathbf{inter}}^{\mathbf{local}}$, $\mathbf{F}_{\mathbf{inter}}^{\mathbf{global}}$): refers to inter-hand features, which capture the broader interactions between two hands, and are intermediate outputs of LGBlock.

In the Method section, we introduce local interaction features and global interaction features, both of which are interaction features, and the latter contains more semantic information about hand interactions than the former.

**Local interaction features** ($\mathbf{F}_{\mathbf{inter}}^{\mathbf{local}}$): obtained by fusing the features of each hand ($\mathbf{F_R}$, $\mathbf{F_L}$) in the Local Unit. It captures the initial interaction information between hands and reflects the relationship between hands in the local range.

**Global interaction features** ($\mathbf{F}_{\mathbf{inter}}^{\mathbf{global}}$): obtained by fusing the local interaction features ($\mathbf{F}_{\mathbf{inter}}^{\mathbf{global}}$) with the features of each hand ($\mathbf{F_R}$, $\mathbf{F_L}$) in the Global Unit. It contains broader and deeper information about the interaction between the two hands, reflecting the complex interaction relationship between the two hands in the global range.

(3) **Global features**(see main manuscript Fig.3: $\mathbf{F}_{\mathbf{R}}^{*}$, $\mathbf{F}_{\mathbf{L}}^{*}$): obtained by adapting the global interaction features ($\mathbf{F}_{\mathbf{inter}}^{\mathbf{global}}$) to each hand ($\mathbf{F_R}$, $\mathbf{F_L}$), which is the final output of the LGBlock, fusing the intra- and inter-hand interaction features to provide overall spatial relationship and interaction information between the two hands.

*Corresponding author: wml@nwsuaf.edu.cn

[1]College of Information Engineering, Northwest A&F University, Yangling 712100, China

[2]Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture, Yangling 712100, China

[3]Shaanxi Key Laboratory of Agricultural Information Perception and Intelligent Service,Yangling 712100, China

‡ These authors contributed equally to this work.

## II. Detailed architecture of LGNet

In the main manuscript, we introduce LGNet, a local and global feature adaptive network for 3D interactive hand mesh reconstruction. It decouples the hand mesh reconstruction process into three stages (joint stage, mesh stage, and refine stage), which can efficiently model the two-hand interaction context, enable high-quality fingertip-level mesh image alignment, and improve the reconstruction accuracy of closely interacting poses to support robotic and VR/AR applications. In this section, we provide a detailed explanation of **the joint stage** as well as **the mesh stage**.

### A. Joint Stage

*1) Local-and-Global feature adaptive block (LGBlock) :*

**Local unit: Graph Convolution for Two-Hand Modeling.** Based on [1], [2], [3], we use the fully-connected (FC) layer $\mathrm{FC}(\cdot)$ to convert $\mathbf{F}$ into a more compact feature vector $\mathrm{FC}(\mathbf{F})$, which is shared among all vertices. We then combine the dense matching encoding (positional embedding) $\mathrm{d_i}$ of the $\mathrm{i}^{\mathrm{th}}$ vertex with the shared vectors to generate the feature $\mathbf{F}_{\mathbf{V}}^{\mathbf{i}}$ for each vertex.

By stacking $\mathbf{F}_{\mathbf{V}}^{\mathbf{i}}$, we obtain $\mathbf{F}_{\mathbf{V}}^{\mathbf{t}} \in \mathbb{R}^{N \times f}$, where $\mathrm{t} = 0$. Subsequently, based on [1], [2], [3], we perform Chebyshev spectral graph CNN operations at each t-th ($\mathrm{t} = 0, 1, 2$) block to convert the input vertex features $\mathbf{F}_{\mathbf{V}}^{\mathbf{t}}$ into $\mathbf{F}_{\mathbf{GCN}}^{\mathbf{t}}$. This operation effectively utilizes the local structural information of the graphical data, and gradually extracts and fuses higher-level local interaction feature representations through multi-layer stacking and graph convolution operations, so that the models are all capable of obtaining richer information from local neighborhoods and encoding this information into higher-level features to capture the complexity and abstract features of the hand data. This operation can better reflect the relationship between hand gestures and movements, thus improving the performance and generalization ability of the models.

**Local unit: Local-and-Global Feature Adaptive Module(LGFA).** In Algorithm 1, we show the inference process of LGFA in the Local unit of LGBlock. LGFA is a core component of LGBlock that converts left and right hand features into new tokens and passes the converted tokens to the Transformer. The LGFA module enables our system to 1) be more robust to similarities between two hands by separating left and right hand features, and 2) avoid the hand feature ambiguity problem and correctly compute correlations between input tokens. It innovatively converts left and right hand features into two new tokens, FCToken and SAToken, as input tokens.

We extract $\mathbf{q}$ from the FCToken because the first and second halves of the FCToken mainly contain information about the two hands, which makes it possible to capture useful interactions between the two hands. The FCToken $\mathbf{t_f} \in \mathbb{R}^{hw \times c}$ is obtained by passing the two hand features ($\mathbf{F_R}$ and $\mathbf{F_L}$) to a fully connected layer is obtained. Before passing the two hand features, we connect and reshape the two hand features ($\mathbf{F_R}$ and $\mathbf{F_L}$) from $\mathbb{R}^{h \times w \times 2c}$ to $\mathbb{R}^{hw \times 2c}$.

We extract $\mathbf{k}$ and $\mathbf{v}$ from the SAToken because the structure of the SAToken causes it to be unable to capture the interaction between the two hands, but only information about the right and left hands.The SAToken $\mathbf{t_s}$ is obtained by passing the two hand features ($\mathbf{F_R}$ and $\mathbf{F_L}$) to a Self-Attention (SA) transformer [4].To this end, we first reshape the dimensions of $\mathbf{F_R}$ and $\mathbf{F_L}$ from $\mathbb{R}^{h \times w \times c}$ to $\mathbb{R}^{hw \times c}$. Then, we connect the reshaped $\mathbf{F_R}$ and $\mathbf{F_L}$ with the class token $\mathbf{t}_{\text{cls}} \in \mathbb{R}^{1 \times c}$. The class token $\mathbf{t}_{\text{cls}}$ is used as a learnable one-dimensional embedding vector to learn general two-handed information [4], [5], [6]. We use $\mathbf{t}_{\text{con}} \in \mathbb{R}^{l \times c}$ to denote the connection marker, where $l = 2hw + 1$. Then, we extract the query $\mathbf{q}_{\text{SA}}$, key $\mathbf{k}_{\text{SA}}$ and value $\mathbf{v}_{\text{SA}}$ from the connection token $\mathbf{t}_{\text{con}}$ using a separate linear layer. The dimensions of $\mathbf{q}_{\text{SA}}$, $\mathbf{k}_{\text{SA}}$ and $\mathbf{v}_{\text{SA}}$ are the same as those of $\mathbf{t}_{\text{con}}$.

**Global unit: Hand Global Interaction Feature Adaptation.** In Algorithm 2. We show the inference process for LGBlock. We use a fully connected layer to reduce the channel dimension of the adaptive global interaction features($\mathbf{F}^{\mathbf{global}}_{\mathbf{inter}}$) from $c$ to $c/4$.Eventually, we connect the adaptive global interaction features and their corresponding hand features ( $\mathbf{F_R}$ or $\mathbf{F_L}$ ) along the channel dimension, which become the final features for each hand, denoted by $\mathbf{F}^*_{\mathbf{R}}$ and $\mathbf{F}^*_{\mathbf{L}}$.

We observe that using $\mathbf{F}^{\mathbf{global}}_{\mathbf{inter}}$ can mitigate the ambiguity due to self-similarity between hands, showing that global interaction information provides a powerful disambiguation cue for local visual features. When there is significant self-occlusion between hands, global interaction information can enhance local visual features in the invisible region, significantly improving the robustness of the estimation of occluded hands. We also observe that using global interaction information can help the network focus on regions that are easily overlooked, such as the dark region.

*2) Joint feature extractor :* The joint feature extractor will extract the 2.5D joint coordinates ($\mathbf{J_R}$ or $\mathbf{J_L}$) and joint features ($\mathbf{F_{JR}}$ or $\mathbf{F_{JL}}$) from the adaptive hand features ($\mathbf{F}^*_{\mathbf{R}}$ or $\mathbf{F}^*_{\mathbf{L}}$). The joint feature extractor [7], [6] efficiently extracts the joint features guided by the estimated 2.5D joint coordinates. The joint features not only contain local information of each hand joint, but also imply global contextual information, which is crucial for the 3D hand mesh reconstruction. This method of integrating local and global information helps to improve the accuracy and robustness of hand movement recognition, thus providing a more reliable basis for hand movement analysis.

**Algorithm 1** algorithm of LGFA in a PyTorch-style

```
class LGFA(nn.Module):
    def __init__():
        FC=nn.Linear(1024,512)
        clstoken=nn.Parameter(512,1)
        Q_SA = nn.Linear(512, 512)
        K_SA = nn.Linear(512, 512)
        V_SA = nn.Linear(512, 512)
        #reshape
        φ:ℝ^{1024×8×8} → ℝ^{64×1024}
        ϕ:ℝ^{512×8×8} → ℝ^{512×64}
        η:ℝ^{512×64} → ℝ^{129×512}
        softmax=nn.Softmax(dim=-1)
        MLP=nn.Sequential(
                nn.Linear(512,512*4),
                nn.Linear(512*4,512))
        d_kSA=512
        d_k=512
    def forward(F_R,F_L):#F_R and F_L ∈ ℝ^{512×8×8}
        #get FCtoken
        t_f=FC(φ(torch.cat(F_R,F_L)))#∈ ℝ^{64×512}
        #get SAtoken
        F_R=ϕ(F_R)#∈ ℝ^{512×64}
        F_L=ϕ(F_L)#∈ ℝ^{512×64}
        t=η(torch.cat(t_cls,F_R,F_L))∈ ℝ^{129×512}
        q_SA,k_SA,v_SA=Q_SA(t),K_SA(t),V_SA(t)
        a=softmax((q_SA k_SA^T)/√d_kSA)v_SA + t
        t_s=MLP(a)+a
        #process two tokens
        q=Q_SA(t_f)
        k=K_SA(t_s)
        v=V_SA(t_s)
        a'=softmax((qk^T)/√d_k)v + t_f
        F^local_inter=a'+MLP(a')
return F^local_inter
```

### *B. Mesh Stage*

In the mesh stage, for each hand, the Self-joint Transformer (SJT) is a standard self-attention transformer. For each hand, it enhances the joint features ($\mathbf{F_{JR}}$ or $\mathbf{F_{JL}}$) [4], [6] by self-attention and thus outputs ($\mathbf{F}^*_{\mathbf{JR}}$ or $\mathbf{F}^*_{\mathbf{JL}}$). We find that SJT can implicitly take into account the kinematic structure of the hand joints through self-attention, and thus is useful for enhancing joint features.

Rough meshes are recovered from low-resolution global features, many local features are lost as a result, and the resulting mesh is mostly smooth with undersampled boundaries. As a result, the rough mesh may not align well with the user's hand in the input image. However, this stage can quickly generate a 3D rough hand mesh that captures the overall shape of the hand $\mathbf{M_r}$. Subsequently, the refine stage allows fine-tuning of the mesh by regressing the offset vectors of each vertex, which are small and easy to learn. This two-stage mesh generation and fine-tuning process ensures that we are able to fine-tune the hand mesh to better

**Algorithm 2** algorithm of LGBlock in a PyTorch-style

```
class LGBlock(nn.Module):
    def __init__():
        LGFA_local=LGFA()
        LGFA_R=LGFA()
        LGFA_L=LGFA()
        LGFA_global=LGFA()
        FC=nn.Linear(512,128)
    def forward(F_R,F_L):#F_R and F_L ∈ R^{512×8×8}
        #Local unit:local interaction feature extract
        F_inter^local=LGFA_local(F_R,F_L)
        #Global unit:Global interaction feature adaptation
        F_interR=LGFA_R(F_R,F_inter^local)
        F_interL=LGFA_L(F_L,F_inter^local)
        F_inter^global=LGFA_global(F_interR,F_interL)
        F_R^* = Concat(F_R, FC(F_inter^global))
        F_L^* = Concat(F_L, FC(F_inter^global))
return F_R^*,F_L^*
```

match the hand shape in the input image while preserving the overall shape. This design of the refinement stage allows our method to achieve more accurate results in real-world applications, especially for application scenes that require high-precision hand alignment, such as augmented reality and virtual reality.

## III. Detailed experiments

### A. Datasets and evaluation metrics

**InterHand2.6M dataset.** InterHand2.6M [9] is the first and only publicly available dataset for two-hand interaction, providing multi-view RGB images with two-hand mesh and joint 3D annotation. Instead of using multi-view information, we treat all images as single-view images. This dataset is highly challenging, containing complex two-hand interaction poses and covering large-scale perspective changes. This large-scale real-captured dataset includes a total of 1,361,062 frames for training, 849,160 frames for testing, and 380,125 frames for validation. It offers accurate human (H) and machine (M) 3D pose and mesh annotations. The dataset is divided into two subsets: interacting hands (IH) and single hand (SH). We trained and evaluated LGNet on the 5 FPS IH subset with H+M annotations, discarding invalid annotations based on the hand type valid annotation provided by [9]. Ultimately, 366K training samples and 261K testing samples were utilized from InterHand2.6M.During preprocessing, we crop the hand region based on the 2D projection of hand vertices and resize it to a resolution of $256 \times 256$.

**HIC dataset.** To demonstrate the generalization ability of our proposed LGNet, we further presented results on the HIC [10] dataset. Unlike the InterHand2.6M [9] dataset, which has uniform backgrounds and lighting sources, the HIC dataset includes more diverse backgrounds and natural lighting. The HIC dataset was only used for evaluation.

**In-the-wild Datasets.** We conducted qualitative experiments on the RGB2Hands dataset [11] and the EgoHands dataset [12]. The RGB2Hands dataset contains 4 sequences of videos covering different types of two-hand interactions, while the EgoHands dataset consists of 48 egocentric videos capturing complex two-person interactions such as playing chess. These datasets contain complex interacting hand samples, diverse backgrounds, realistic lighting conditions, and varying image quality, which enable a comprehensive evaluation of the generalization ability of our method.

**Evaluation metrics.** For fair comparison, we follow prior research [9], [8], using the wrist as the root joint for joint alignment and scaling the predicted results according to the ground-truth bone length during evaluation. Particularly, in qualitative experiments, we do not perform root joint alignment and scaling.

### B. Comparisons with state-of-the-art methods

Following previous works [9], [13], we measured MPVPE after scaling alignment on meshes with ground truths (GTs). We re-evaluated these methods in the same manner, using the true positions of wrist joints for hand alignment and scaling the estimated results according to the ratio of predicted bone length to the ground truth bone length. Specifically, we trained our model according to the official split and reported results on the official test set of the InterHand2.6M dataset for fair comparison.

Furthermore, we introduce a lighter version of LGNet (Light) for fair comparison with the less parameterized IntagHand [3]. Even with reduced channel dimension to $c = C/8$, the performance of LGNet (Light) surpasses that of IntagHand [3], further demonstrating the effectiveness of our proposed approach.

We attribute this success to the dense mesh reasoning capability of GCN, our novel attention-based module LGBlock, and our three-stage mesh reconstruction model structure. GCN exhibits excellent reasoning capability on mesh data, effectively extracting both local and global relationships of the hand. Our novel attention-based module LGBlock integrates local features to capture global interaction features of the hand and adapts to each hand individually. Furthermore, our three-stage mesh reconstruction pipeline combines feature extraction, hand joint regression, and mesh refinement processes organically, making the entire reconstruction process more effective and reliable.

### C. Additional qualitative comparisons

As shown in Figure 8 (the main manuscript), our method can generate high-quality two-hand reconstruction results under severe occlusions and various interaction contexts. For cases involving severe self-occlusion and fine-grained interactions (**row1**), our method achieves accurate mesh-image alignment and correctly interprets the relationship between the two hands. Additionally, in instances of severe self-occlusion, where there are few observable pixels for the hand (**row2**) or when the observable region is dark (**row3**), our method accurately reconstructs the corresponding areas. Our

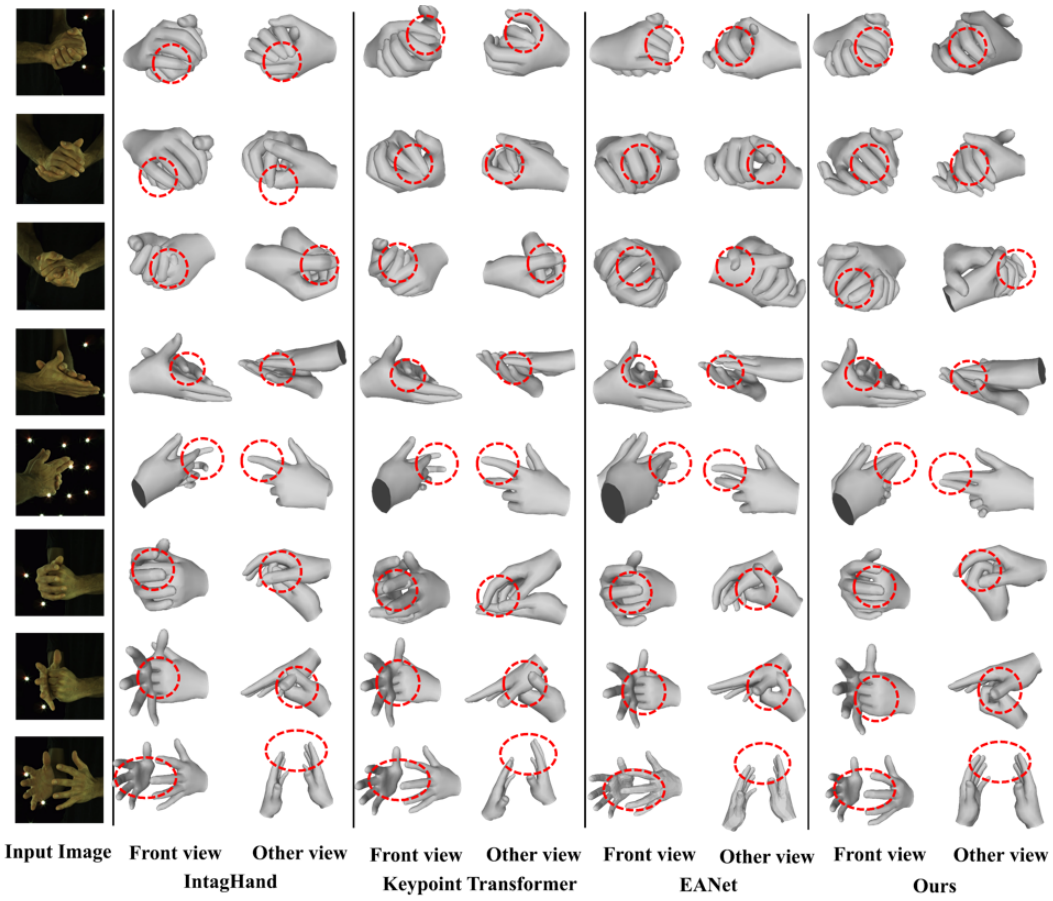


Fig. 1. **Visual comparison with the state-of-the-art method [3], [8], [6] on the InterHand2.6M dataset [9].** The red circles highlight regions where our LGNet is correct, while other methods are wrong. Our method produces more accurate hand poses, while the other methods [6] generate more collisions and inaccurately estimate the relative depth between the left and right hands.

method may generate high-quality two-hand reconstruction results even when multiple challenging conditions such as self-occlusion, tight interaction, blurring, or shadowing occur simultaneously or in combination.

As shown in Figure 9 (the main manuscript), compared to the previous state-of-the-art method, our method produces more realistic finger interactions and fewer collisions between the two hands. Our method can avoid the collapse of estimated meshes and better avoid unreasonable intersections between hands (**row1**). This indicates that the LGBlock module contributes to capturing the spatial relationship between the hands in our model. Additionally, our method achieves better alignment between the mesh and the image (**row2**), and it can still infer the pose of occluded hands based on fine-grained visual cues and global information. The results demonstrate the superior performance of our LGNet in accurately estimating hand poses and interactions.The results demonstrate the superior performance of our LGNet in accurate hand pose and interaction estimation.

As shown in Figure 10 (the main manuscript), we demonstrate the generalization ability of our method on in-the-wild images. our method performs well on real data captured by a common USB camera (**rows 1-3**). Additionally, without additional training, our model achieves excellent results on the RGB2Hands dataset (**rows 4-7**) and EgoHands dataset (**rows 8-11**), showing strong generalization across different viewpoints such as third-person or egocentric perspectives. Moreover, as indicated by rows 1 and the last row, our method is more robust to different backgrounds and lighting conditions. Due to its ability to utilize global information to enhance visual features, our method exhibits relative robustness to object perturbations and maintains hand structure effectively, showing significant advantages in preserving hand structure.Furthermore, our model achieves an inference speed of 30fps on a single NVIDIA RTX 3090 GPU, enabling potential real-time applications in the future.

In the below figures, we further show qualitative comparisons with other state-of-the-art methods[3], [8], [6]. Specifically, in Figure 1, we show visual comparisons on the InterHand2.6M [9] dataset. In Figure 2, we demonstrate the generalization ability of our method on in-the-wild images.

### *D. Ablation study*

**Effects of decoupled design.** We conducted an ablation study, evaluating three cases as follows: 1) remove the joint stage. In this case, we used fully connected layers to regress hand joints without going through the joint stage processing. 2) remove the mesh stage. In this case, we

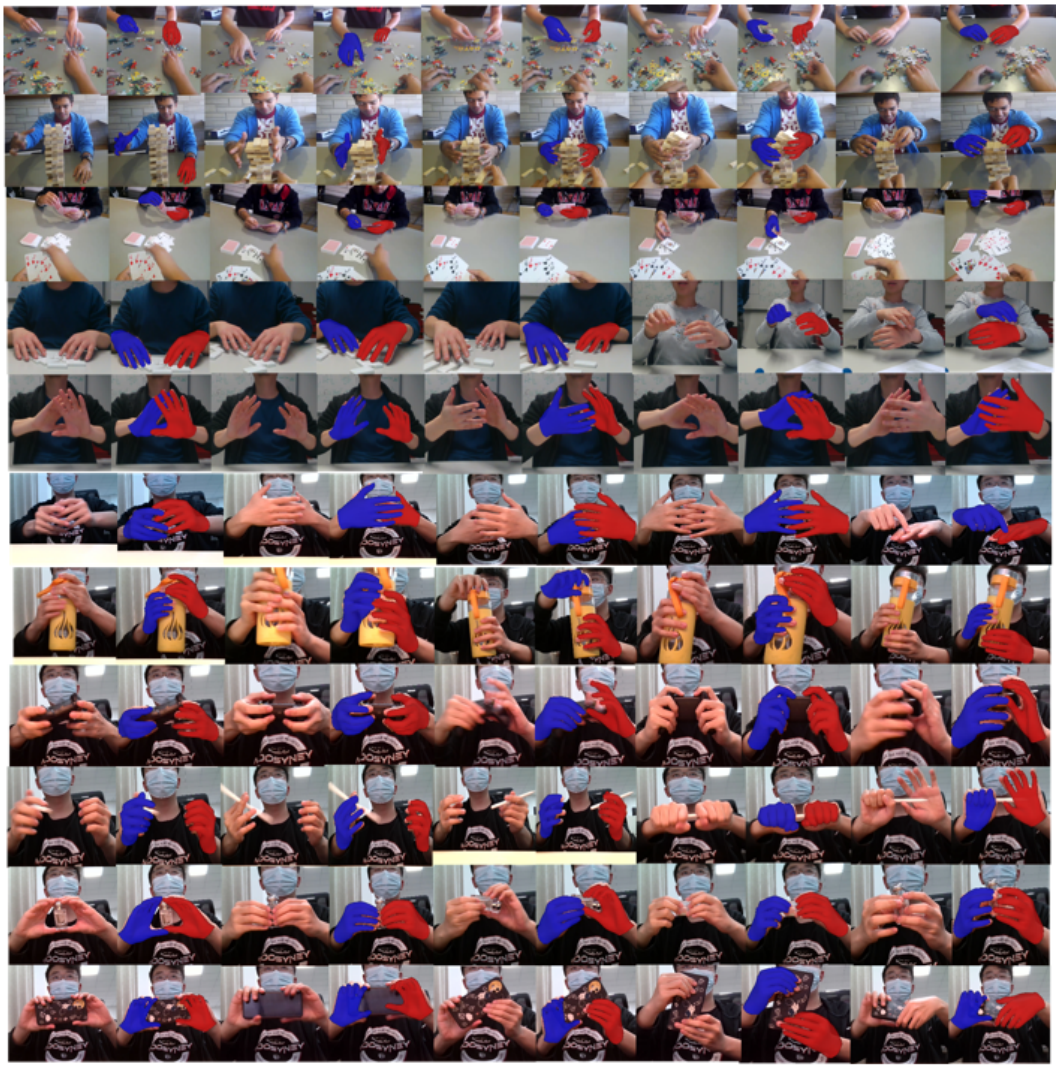

Fig. 2. **Qualitative results on in-the-wild images.** In each part, the left is the input image, and the right is the result predicted by our method.

used fully connected layers to regress MANO parameters to generate the final hand mesh without going through the mesh stage processing. 3)remove the refine stage. In this case, we treated Mr as the final hand mesh without going through the refinement stage processing. The ablation results are reported in the first three rows of Table 3 (the main manuscript). We observed significant performance drops when removing any stage, indicating the effectiveness of each stage and their contributions to the overall framework performance. Additionally, please refer to Figure 11 (the main manuscript) for visual comparison.

**Effects of joint stage & loss terms.** In addition, we evaluated the contributions of each component and loss term in the joint stage for the following cases: 1) not separate left and right hand features; 2) remove the LGBlock; 3) remove the Local unit; 4) remove the Global unit; 5) remove the $L_{normal}$; 6) remove the $L_{edge}$; 7) remove the $L_{lap}$. The ablation results are reported in Table 3 (the main manuscript). Clearly, our complete pipeline performs the best across all metrics, and removing any component leads to a decrease in overall performance, indicating that each component contributes to improving the final result. These results further validate the effectiveness of each component in our method and highlight their importance within the entire framework.

## IV. Discussion

### A. Societal impacts

Our LGNet enables robust 3D interacting hand mesh reconstruction, which is critical for advancing robotics applications that require precise hand interaction modeling. This technology has significant societal impacts, particularly in the fields of robotics, augmented reality (AR) and virtual reality (VR), where it can bridge the gap between the physical and virtual worlds. The ability to accurately capture and reconstruct dynamic interactions between two hands is essential for enhancing the realism and functionality of robotic systems used in teleoperation, virtual collaboration, and remote-controlled robotic surgery. By improving the accuracy and fidelity of these interactions, our work contributes

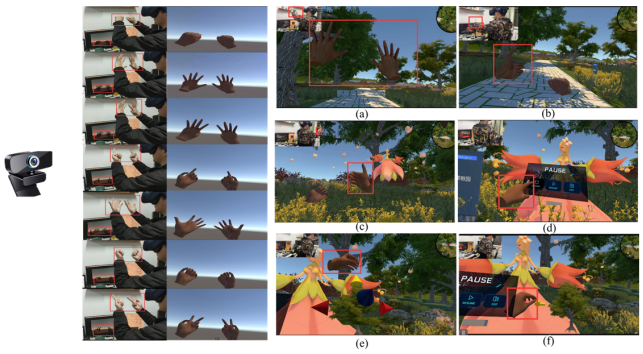


Fig. 3. User operation flow, hand real-time reconstruction effect and software gesture control effect demonstration: (a) Hand modeling demonstration (b) Gesture control roaming. Defined as the left thumb straight, the rest of the four fingers bent, and the palm of the hand toward the camera, you can move forward (c) Gesture to summon the elf, defined as the right thumb straight, the rest of the four fingers bent, and the palm of the hand toward the camera (d) Interacting with the display board (e) Gesture to control the rotation of the plant model (f) Gesture to control the rotation of the plant model.

to the development of more effective and intuitive robotic interfaces, ultimately benefiting industries such as healthcare, manufacturing, and education.

### B. Applications

Based on the algorithm in this paper, a real-time gesture interaction virtual roaming software based on monocular camera is developed by combining virtual reality technology with real life, taking the wetland ecological park as the research object. The software only uses the monocular camera to capture hand information, and transmits the hand position information to the Unity engine through the UDP protocol, which realizes the functions of hand modeling, gesture-controlled roaming, gesture summoning, gesture control and spreadsheet, as well as model moving and rotating. The software realizes real-time rendering and immersive interactive experience, providing users with an immersive virtual roaming experience.

In order to achieve real-time hand capture and control, this system utilizes a monocular camera to capture the joint point information of the hand and perform hand modeling, the generated hand model achieves high-quality real-time gesture reconstruction results, which are consistent with the expected results. Through gesture control, the user can realize roaming in the virtual scene, gesture summoning, gesture control of the exhibition board and control the movement and rotation of the model as shown in Fig Figure 3 .

Effective way to combine new computer virtual technology and ecological protection, with a sense of realism model, only through the monocular camera to capture gestures, and add UI for operation, get rid of heavy equipment, reduce costs. Can let the user have a more immersive experience, in this whole experience process, not only can let people relax to relieve stress, appreciate the interest of the garden art, but also let people feel the happiness of green life, so as to stimulate a positive attitude towards life, enhance the importance of environmental protection, and help to promote the construction of ecological civilization. We demonstrate the feasibility of our algorithm in future applications.

### C. Future Work:

In future research, we plan to design an adaptive refinement stage that focuses on the accurate alignment of boundary vertices to further enhance the representation of 3D hand meshes in robotic tasks. Through the optimisation of this stage, we not only hope to significantly improve the model inference efficiency during complex hand pose reconstruction, but also aim to reduce the computational resource consumption to meet the stringent requirements of real-time and computational efficiency in robotic systems. This improvement is especially critical for achieving fast and accurate hand pose sensing on resource-limited robot platforms.

In addition, we will explore the feasibility of deploying LGNet to mobile robots and smart devices. This work will allow our approach to show great potential for application on portable robotics platforms, especially in mobile scenarios that require real-time human-robot interaction. By enabling efficient 3D hand gesture sensing on mobile devices, we can advance the application of autonomous robots in dynamic environments and lay the foundation for future mobile robots and smart devices in areas such as human-robot collaboration and gesture control.

### D. License of the Used Assets

InterHand2.6M dataset [9] is CC-BY-NC 4.0 licensed.
HIC dataset [10] is publicly available dataset.
RGB2Hands dataset [11] is publicly available dataset.
EgoHands dataset [12] is publicly available dataset.
InterNet [9] are released for academic research only, and it is free to researchers from educational or research institutes for non-commercial purposes.
DIGIT[14] are released for academic research only, and it is free to researchers from educational or research institutes for non-commercial purposes.
InterShape[13] are released for academic research only, and it is free to researchers from educational or research institutes for non-commercial purposes.
Keypoint Transformer[8] are released for academic research only, and it is free to researchers from educational or research institutes for non-commercial purposes.
IntagHand[3] are released for academic research only, and it is free to researchers from educational or research institutes for non-commercial purposes.
DIR[15] are released for academic research only, and it is free to researchers from educational or research institutes for non-commercial purposes.
ACR[16] are released for academic research only, and it is free to researchers from educational or research institutes for non-commercial purposes.

## REFERENCES

[1] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in neural information processing systems*, vol. 29, 2016.

[2] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan, "3d hand shape and pose estimation from a single rgb image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 833–10 842.

[3] M. Li, L. An, H. Zhang, L. Wu, F. Chen, T. Yu, and Y. Liu, "Interacting attention graph for single image two-hand reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2761–2770.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] J. Park, D. S. Jung, G. Moon, and K. M. Lee, "Extract-and-adaptation network for 3d interacting hand mesh recovery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4200–4209.

[7] G. Moon, H. Choi, and K. M. Lee, "Accurate 3d hand pose estimation for whole-body 3d human mesh estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2308–2317.

[8] S. Hampali, S. D. Sarkar, M. Rad, and V. Lepetit, "Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 090–11 100.

[9] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee, "Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 548–564.

[10] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 89–98.

[11] J. Wang, F. Mueller, F. Bernard, S. Sorli, O. Sotnychenko, N. Qian, M. A. Otaduy, D. Casas, and C. Theobalt, "Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video," *ACM Transactions on Graphics (ToG)*, vol. 39, no. 6, pp. 1–16, 2020.

[12] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1949–1957.

[13] B. Zhang, Y. Wang, X. Deng, Y. Zhang, P. Tan, C. Ma, and H. Wang, "Interacting two-hand 3d pose and shape reconstruction from single color image," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 354–11 363.

[14] Z. Fan, A. Spurr, M. Kocabas, S. Tang, M. J. Black, and O. Hilliges, "Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 1–10.

[15] P. Ren, C. Wen, X. Zheng, Z. Xue, H. Sun, Q. Qi, J. Wang, and J. Liao, "Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8014–8025.

[16] Z. Yu, S. Huang, C. Fang, T. P. Breckon, and J. Wang, "Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 955–12 964.