IFVONet: Integrating Inter-Frame Variation and Occlusion Awareness for 3D Hand Mesh Reconstruction

Shiyao Wu
School of Computing and Mathematical Sciences
University of Leicester
Leicester, UK
Email: sw658@leicester.ac.uk

John Panneerselvam
Department of Computer Science
University of Exeter
Exeter, UK
Email: j.panneerselvam@exeter.ac.uk

Lu Liu *
Department of Computer Science
University of Exeter
Exeter, UK
Email: l.liu3@exeter.ac.uk

Rajeev Raman
School of Computing and Mathematical Sciences
University of Leicester
Leicester, UK
Email: rr29@leicester.ac.uk

Tianjin Huang
Department of Computer Science
University of Exeter
Exeter, UK
Email: t.huang2@exeter.ac.uk

***Abstract*—Reconstructing 3D hand meshes from video files is significantly challenging due to objects in the video often occluding the hand during manipulation. These occlusions can greatly reduce the quality of information extracted from the obscured regions and decrease temporal hand coherence over time. Existing approaches focus primarily on global occlusion regions but overlook temporal hand coherence, which limits their performance. Herein, we propose a novel framework called IFVONet, designed to improve 3D hand mesh reconstruction by effectively capturing inter-frame variations and improving the recovery of global occlusions. IFVONet comprises three key components: (1) Pixel-Domain Variation Module for identifying inter-frame variations across adjacent frames, enhancing temporal hand coherence. (2) Enhanced Global Occlusion Recovery Module for integrating hand information into global occlusion representation, thereby improving the accuracy of occlusion feature recovery. (3) Hand Regression Module for dynamically aggregating hand information from inter-frame variations and globally recovered occlusion features into comprehensive hand representations, ultimately leading to enhanced 3D hand reconstruction. Extensive experiments on the HO3D-v2 and HO3D-v3 datasets demonstrate that our proposed IFVONet achieves state-of-the-art performance on both 3D hand mesh reconstruction and pose estimation.**

***Keywords*-3D Hand Mesh Reconstruction; Transformer; Deep Learning; Video Understanding;**

I. Introduction

Reconstructing 3D hand meshes from RGB frames is crucial for numerous real-world applications, including augmented reality (AR) [1] and behavior understanding [2]. To support these applications, ensuring an optimal user experience is crucial. Therefore, the reconstruction process must be not only accurate and robust but also temporal coherence through time. Recently, several techniques have been proposed for 3D hand mesh reconstruction from a single RGB frame [3], [4], [5], [6]. To extract enough information and alleviate occlusion situations for targeted tasks, some of the existing works utilized spatial attention mechanisms to recover occluded hand feature [3], [7], others proposed texture or lightning models to get a better hand mesh result [4], [8]. These methods are more efficient on single static hand images. However, existing methods suffer severe performance degradation under sequence data due to limited information for a single RGB frame, leading to temporal hand incoherence through time and inaccurate hand pose reconstruction results.

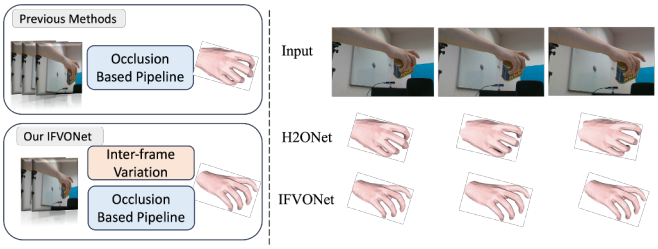

Figure 1. Structural and reconstruction comparison with state-of-art methods, the proposed method IFVONet not only keeps hand coherence and natural hand reconstruction results through time for the reconstruction result but considers inter-frame variation for the structural design.

On the other hand, multi-frame RGB-based approaches have gained attention, which can explore additional temporal information, such as hand motion, to enhance model performance. S^2HAND(V) [9] constrains smooth hand motion by presenting a motion-aware joint loss function to help train a frame-wise model with promising results. SeqHand [10] utilizes ConV-LSTM to extract temporal movement information over consecutive frames. Besides, Inter-frame loss is used by Liu [11] to leverage spatial-temporal consistency for adjacent frames. H2ONet [12] additionally inputs short

* Corresponding Author

2993-396X/24/$31.00 ©2024 IEEE
DOI 10.1109/SWC62898.2024.00120

and long historical frame information to fetch non-occlusion and infuse finger-level information. Even though they can get additional information from adjacent frames, they lack a mechanism to handle the inter-frame variation, leading to incoherent temporal results through time, as shown in Fig.1.

In this paper, We propose IFVONet, an Inter-Frame-Variation and Occlusion-aware Network, to enhance reconstructing 3D by effectively capturing inter-frame variations and improving the recovery of global occlusions. Firstly, to address variations across frames, we incorporate inter-frame interactions at the pixel level into 3D hand reconstruction using an attention mechanism. Secondly, to improve the recovery of global occlusions, we refine the components of the Feature Injection Transformer (FIT) by systematically analyzing the effectiveness of each component on 3D hand reconstruction. Finally, we dynamically fuse the information on inter-frame variations and occlusion recovery using a learned probability map. Our proposed IFVONet framework improves occlusion recovery features and preserves temporal hand coherence by capturing inter-frame variations, leading to enhanced 3D hand reconstruction.

To summarize, our main contributions are as follows:

- We explicitly integrate inter-frame variations into the 3D hand reconstruction process, resulting in enhanced and temporally coherent 3D hand reconstruction.
- We propose a novel framework, IFVONet, which comprises three key modules: a Pixel-Domain module for capturing inter-frame variations, a refined Global Occlusion Recovery module for retrieving occlusion information, and a Variation-Occlusion-Aware hand regression module for fusing inter-frame variation features with recovered global occlusion information.
- Through comprehensive evaluations across two datasets containing HO3D-v2 and HO3D-v3, our results consistently demonstrate IFVONet's significant potential in enhancing the quality and temporally coherence of 3D hand reconstruction, achieving state-of-the-art performance.

The remainder of this paper is organised as follows: Section 2 reviews related work in 3D hand mesh reconstruction from RGB frames along with occlusion-aware pose estimation methods. Section 3 details our proposed IFVONet framework. Section 4 evaluates the performance of our proposed IFVONet framework and discusses the results and section 5 concludes this paper.

II. Related Work

A. 3D hand mesh reconstruction from RGB image

Existing works on 3D hand mesh reconstruction techniques can be divided into two groups as RGB and depth images, according to the type of image they input.

For RGB-based methods, some works estimate the required parameters for MANO[13] to get hand vertices and 3D hand joints [3], [6], [5], [14]. Given the RGB input, Pavlakos1[14] regress pose, shape and camera parameter based on Transformer, which is a representative work for this pipeline, similar to HMR[15] in human mesh reconstruction task. However, others regress hand vertices directly from deep learning networks [16], [17], [18]. Specifically, Jiang [8] utilized the attention-based module to formulate the correlation across vertex and joint features and further regress the coordinates of the vertex.

For depth-based methods, Korrawe et al. [19] started to utilize interactive optimization to deform the hand mesh. Later on, deep learning-based approaches [20] were explored with depth-based methods to improve model performance, such as CNN.

When the hand is occluded by itself or an object in the sequence data, the majority of the current methods produce undesirable temporal hand coherence results, as they largely ignore temporal information.

B. 3D hand mesh reconstruction from RGB video

Some recent efforts have utilised multi-frame information for 3D hand mesh reconstruction in video-based approaches. Compared with the original RGB-based S^2Hand method [4], the S^2HAND(V) [9] adds a quaternion loss function, which can explore joint rotation representation across multi frames. An optical-flow-guided approach is used by Hasson *et* al [21] to take advantage of photometric consistency across time. H2ONet [12] is the most relevant and closely related approach, which explores non-occlusion information by estimating finger-level occlusions and fusing hand-level information from long and short historical frames. However, the existing approaches mainly concentrate on occlusion recovery and do not consider inter-frame variation to improve temporal hand coherence, which is addressed in our proposed IFVONet framework.

C. Occlusion-aware pose estimation

There are three primary methods to consider when estimating occlusion-based position, including attention mechanisms, data augmentation and temporal information.

The attention mechanism-based methods utilise spatial information [22], [7], [3], [23]. Zhou *et* al [7] estimated occlusion-aware attention map to address redundant data and recover missing information from occlusion during data cleansing. For hand pose estimation under occlusion, HandOccNet [3] explores the relationship between occlusion and non-occlusion region to obtain complete hand information. Data augmentation methods simulate the occlusion situation [24], [25], [26]. Cheng *et* al [25] masks out key-points during the training phase to simulate low to severe occlusion. Even though these techniques can improve model performance, the generated occlusion is synthetic when compared with real occlusion. Temporal methods adopt temporal information from the sequence file [12], [27], [28]. Cheng *et* al

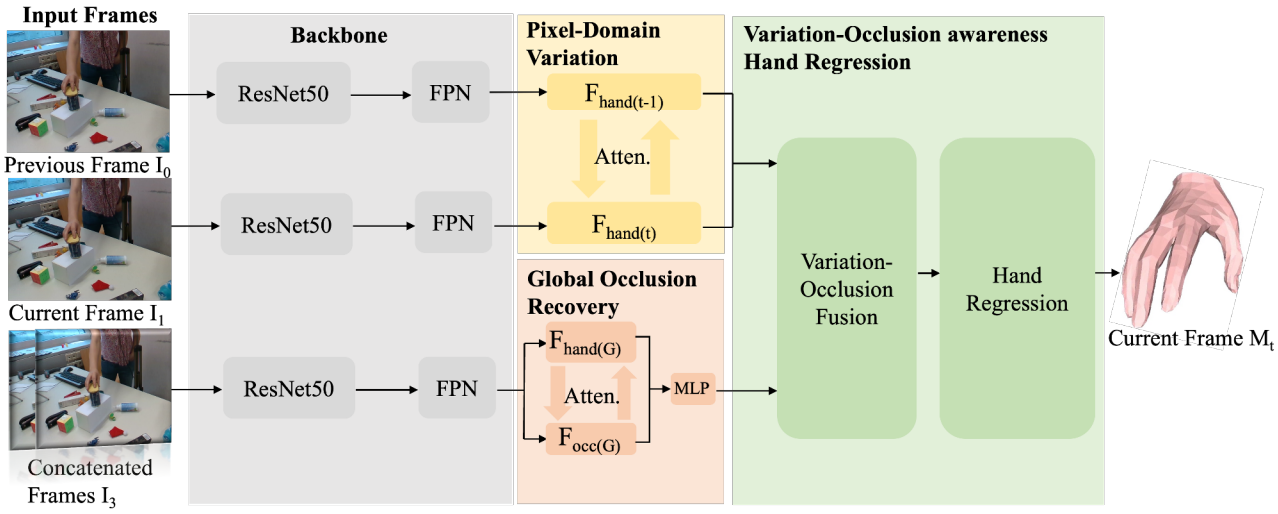

Figure 2. Pipeline for proposed IFVONet. IFVONet contains a three-branch backbone, Pixel-Domain Variation (PDV), Global Occlusion Recovery (GOR), Variation-Occlusion Awareness hand regression. We first extract the global hand feature, global occlusion feature, t frame hand feature and $t-1$ frame hand feature from three-branch backbone. Inter-frame variation is extracted from the proposed Pixel-Domain Variation while keeping the temporal hand coherence. Global Occlusion Recovery is proposed to alleviate the occlusion situation. Variation-Occlusion awareness hand regression generates 3D hand mesh reconstruction, 2D hand pose estimation, and 3D hand pose estimation.

[27] proposed a temporal CNN to complete a 3D pose with occluded 2D key-points. Additionally, H2ONet [12] predicted finger-level occlusion probabilities to guide hand-level feature infusion over short and long time frames.

Different from the above methods, our proposed IFVONet framework embeds temporal information with cross-pixel attention to extract the inter-frame variation on the pixel level, followed by cross-region attention to learn the correlation between occlusion and hand regions with a dynamic fusion of relevant information.

III. IFVONet Framework

Our goals are to estimate 3D hand mesh vertex V_{3D}, 3D hand pose J_{3D} and 2D hand pose J_{2D} for each frame, given two sequenced 2D RGB images $I \in R^{256\times256\times3}$ containing the hand-object interaction information from adjacent frames. To achieve the above goals, the IFVONet framework is proposed in this paper, which consists of three modules within the output from a three-branch backbone: a Pixel-Domain Variation module (PDV), a Global Occlusion Recovery module (GOR) and a Variation-Occlusion awareness hand regression.

Given the adjacent frames, three different inputs are firstly generated, including hand image I_0 for the current frame, hand image I_1 for the previous frame and concatenated images I_3. ResNet50 with FPN is utilised as a backbone to obtain the hand feature for each frame $F_t \in R^{256\times32\times32}$ and $F_{t-1} \in R^{256\times32\times32}$ and global hand feature $F_h \in R^{256\times32\times32}$ and the global occlusion feature $F_o \in R^{256\times32\times32}$ for the concatenated frames.

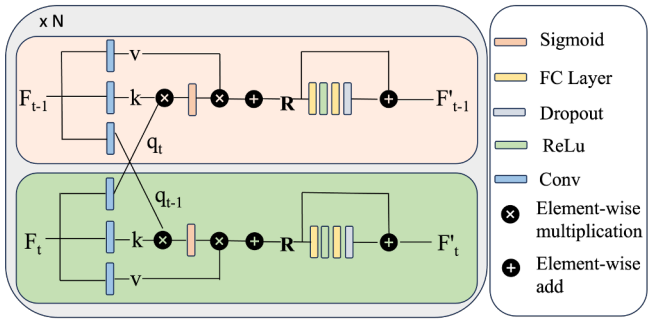

Figure 3. The pipeline for Pixel-Domain Variation (PDV). Hand feature from the current and previous frames to extract the pixel-level variation and keep the unchanged hand region, improving temporal hand coherence through time.

A. Pixel-Domain Variation (PDV)

Temporal hand coherence plays an important role in sequence data, leading to a better user experience in real-world applications. Therefore, Pixel-Domain Variation (PDV) is proposed to distinguish the variation across frames on the pixel level, which can distinguish the inter-frame variation and unchanged hand region to maintain hand pose stability through time. The illustration of PDV is shown in Figure 3. PDV is a Transformer-based module which takes two features, F_{t-1} and F_t. Two same sub-modules are adopted in the PDV to consider the pixel variation across frames. Different from the standard Transformer, a symmetric cross-pixel attention (CPA) module is utilised to formulate this variation.

Three 1x1 Convolution layers are first performed to obtain

Q_T, K_T and V_T ($T \in t-1, t$) indicating the query, key and value feature of each frame. Then the query feature Q_T of one frame to fetch the key feature K_T and the value feature V_T of the other frame are utilised through Multi-Head Attention (MHA) as:

$$F_{t-1\to t} = sigmoid(\frac{Q_{t-1}K_t^T}{\sqrt{d}})V_t$$

$$F_{t\to t-1} = sigmoid(\frac{Q_tK_{t-1}^T}{\sqrt{d}})V_{t-1}$$

where $F_{t-1\to t}$ and $F_{t\to t-1}$ are the cross-pixel attention feature encoding the variation region between two frames, and d is a normalisation constant. Afterwards, the output frame features $F' \in R^{256\times32\times32}$ are obtained by a point-wise MLP layer f_p as:

$$F'_t = f_p(F_{t-1\to t})$$

$$F'_{t-1} = f_p(F_{t\to t-1})$$

B. Global Occlusion Recovery (GOR)

The Global Occlusion Recovery module (GOR), as shown in Fig.4, is improved from the Feature Injecting Transformer (FIT) [3], which contains sigmoid-based and normalisation-based attention modules, to enhance the correlation between hand and occlusion information. In addition, the contribution of the FIT is systematically analysed, including the contribution of its sub-modules to the hand information and occlusion information, thereby paying more attention to its related regions.

The sigmoid-based attention module can filter the low-related region [3]. Moreover, the occlusion feature not only contains the occluded hand information but also contains undesired information(e.g., background). Therefore, the sigmoid-based attention module is utilised to generate the correlation map on the occlusion region rather than the hand region. Query q_{sig} and key k_{sig} are extracted from the F_o and the F_h separately by two 1x1 convolution layer. Therefore, the correlation map C_{sig} can be obtained with:

$$C_{sig} = sigmoid(\frac{q_{sig}k_{sig}^T}{\sqrt{d_{k_{sig}}}})$$

Compared with the traditional attention method, the normalisation attention mechanism offers a simple yet efficient way to improve the model performance [29]. As the hand feature not only contains the information under the visible hand feature but also includes invisible hand information. Therefore, the normalisation-based attention module scores the hand regions, which helps to mitigate the occlusion information on the invisible hand region. Then, query q_{norm} and key k_{norm} are extracted from the F_o and the F_h separately. Afterwards, the module generates the score map S_{norm} as follows:

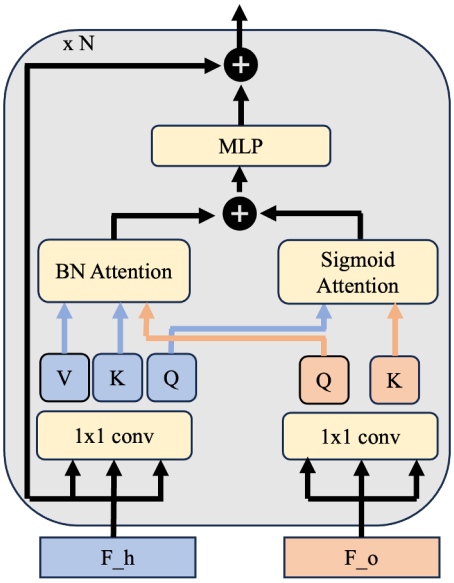

Figure 4. The overall pipeline for Global Occlusion Recovery (GOR). The Transformer Block contains sigmoid-based attention and a Normalization-based attention module, using the addition operation with concatenated F_o into F_h.

$$S_{norm} = BN(f_{1\times1}(\frac{q_{sig}k_{sig}^T}{\sqrt{d_{k_{sig}}}}))$$

Similar to HandOccNet [3], which considers the injection of the hand and occlusion region, instead of applying a multiplication, an addition operation is simply adopted between score map S_{norm} and the correlation map C_{sig}, aiming to obtain the high-related region to address in the missing information:

$$C = C_{sig} + S_{norm}$$

Then, the invisible hand information is obtained by multiplying the correlation map C by value v_h, which is obtained with one 1×1 convolution layer. Therefore, the refined hand feature F'_h is obtained with the MLP module and layer normalisation layer. Besides, as the value v_h preserves the essential information for the hand information, a residual connection between F'_h and v_h is utilised and the output feature $F_{GOR} \in R^{256\times32\times32}$ is obtained by:

$$F_{GOR} = v_h + MLP(LN(C \times v_h))$$

C. Variation-Occlusion awareness Hand Regression

As the detailed inter-frame variation provided by PDV is crucial to maintain temporal hand coherence in the sequence data, the utilisation of variation information is important for global occlusion recovery. Different from the previous fusion, which directly adds or multiplies the output,

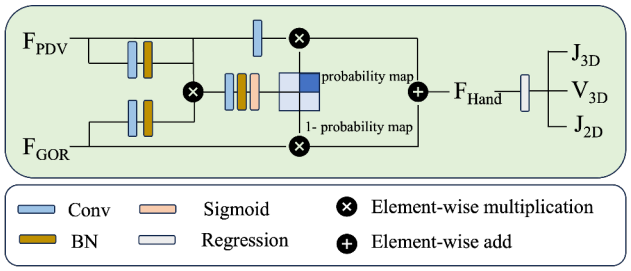

Figure 5. The pipeline for Variation-Occlusion awareness hand regression. The possibility map is calculated to leverage the balance between inter-frame variation and global occlusion information.

Variation-Occlusion Fusion (VOF) is proposed for the global recovered occlusion feature F_{GOR} to selectively learn useful variation features from F_{PDV} without being overwhelmed, which can leverage the information from global recovered **occlusion** feature and inter-frame **variation**, as shown in Fig.5 Specifically, within the output of PDV, the F'_t and F'_{t-1} are firstly concatenated to obtain concatenated variation hand feature F_{PDV} across frames. Furthermore, the F_{PDV} is regarded as the backup, enabling it to provide the required variation information to F_{GOR}. Then, the probability of these two pixels belonging to the hand region is calculated:

$$p = Sigmoid(f_{gor}(F_{GOR}) \times f_{pdv}(F_{PDV}))$$

where f_{gor} and f_{pdv} represent the 1×1 convolution layer following with Batch Normalisation layer. If the value of p is high, we would argue that F_{GOR} contains more rich and accurate hand information, and vice versa. Therefore, the refined hand feature $F_{Hand} \in R^{256\times32\times32}$ can be written as:

$$F_{Hand} = pF_{GOR} + (1-p)f_{1x1}(F_{PDV})$$

Given the variation-occlusion awareness hand feature F_{Hand}, regression head produces 2D joint pose estimation, 3D joint pose estimation and 3D hand mesh vertices based on MANO [13]. The hourglass block first generates sub-heatmaps $H \in R^{256\times32\times32}$ for each joint and 2D hand joint coordinates $J_{2D} \in R^{21\times2}$ within the input of enhanced feature F_{Hand}. Secondly, four residual blocks and two different fully-connected layers first concatenate each sub-heatmap and generate pose $\theta \in R^{48\times3}$ and shape $\beta \in R^{10}$ parameters. Four linear layers are adopted for shape regression and one for pose regression respectively. To obtain 3D hand vertices $V_{3D} \in R^{778\times3}$ and 3D hand joint coordinates $J_{3D} \in R^{21\times3}$, the regressed pose θ parameter multiplies the joint regression matrix and applies forward kinematics based on MANO [13].

D. Loss Functions

To train the proposed IFVONet, we use common loss function for 2D and 3D results, including 3D hand mesh vertices, 2D joint coordinates and 3D joint coordinates. First, $\mathcal{L}_1$ loss calculates the Euclidean distance between the prediction and ground truth, including the 3D mesh vertices loss $\mathcal{L}_{mesh}$, 2D joint coordinates loss $\mathcal{L}_{pose2D}$ and 3D joint coordinates loss $\mathcal{L}_{pose3D}$. Formally, we have

$$\mathcal{L}_{mesh} = \sum_{i=0}^{1} ||V_{t-i} - \hat{V}_{t-i}||_1$$

$$\mathcal{L}_{pose2D} = \sum_{i=0}^{1} ||J^{2D}_{t-i} - \hat{J}^{2D}_{t-i}||_1$$

$$\mathcal{L}_{pose3D} = \sum_{i=0}^{1} ||J^{3D}_{t-i} - \hat{J}^{3D}_{t-i}||_1$$

As the MANO layer is applied for the regression head, to ensure the performance of MANO and avoid outliers for these essential parameters, we use $\mathcal{L}_1$ loss for pose β and shape θ parameters. Formally, we have

$$\mathcal{L}_{mano\_pose} = \sum_{i=0}^{1} ||P_{t-i} - \hat{P}_{t-i}||_1$$

$$\mathcal{L}_{mano\_shape} = \sum_{i=0}^{1} ||S_{t-i} - \hat{S}_{t-i}||_1$$

Our overall loss function is $\mathcal{L}_{total} = \mathcal{L}_{mesh} + \mathcal{L}_{pose2D} + \mathcal{L}_{pose3D} + \mathcal{L}_{mano\_pose} + \mathcal{L}_{mano\_shape}$.

IV. Experiments

A. Experiment Environment

The proposed method is implemented using PyTorch [30]. Adam optimiser [31] with batch size 64 is utilised for our model training on NVIDIA A100 Tensor Core GPU. 40 epochs are taken with an initial learning rate of 10^{-4}, which is divided by 10 at the 20^{th} epoch within the usage of the step learning scheduler. The input images are resized into 256 x 256 and augmented by random scaling and rotating.

B. Dataset

The HO3D-v2 [32] dataset includes 66,034 samples from 55 sequences for training and 11,524 samples from 13 sequences for testing, and contains hand-object interaction scenario on RGB sequences with different levels of occlusion in three categories, including seen object with seen hand, unseen object with seen hand and seen object with unseen hand. 2D joints, MANO-based vertex and 3D hand joints are included in the annotations.

Compared with the HO3D-v2 [32] dataset, the HO3D-v3 [33] dataset provides more image data and more accurate annotations, which have been released recently. This dataset contains 103,462 images for hand-object interaction, which are divided into 83,325 training images and 20,137 testing images with 3D hand annotation.

Table I
RESULTS ON HO3D-V2 DATASET AFTER PA. **BEST** AND SECOND-BEST SCORES.

Approaches	J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑
Hasson et al. [21]	11.4	77.3	11.4	77.3	42.8	93.2
Hasson et al. [34]	11.1	-	11.2	-	46.4	93.9
TempCLR [28]	10.6	-	10.6	-	48.1	93.7
Liu et al. [11]	9.8	-	14.7	81.2	53.0	95.7
HOISDF [35]	9.2	-	-	-	-	-
HandOcc(V) [3]	9.2	81.5	9.3	81.5	54.2	95.9
H2ONet [12]	**9.0**	**82.0**	9.1	81.9	54.7	96.0
Our IFVONet	**9.0**	81.9	**9.0**	**82.0**	**55.4**	**96.2**

Table II
RESULTS ON HO3D-V2 DATASET BEFORE PA. **BEST** AND SECOND-BEST SCORES.

Approaches	J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑
Liu et al. [11]	30	49.0	28.9	50.3	23.2	68.5
H2ONet [12]	26.9	52.4	26.1	53.5	24.9	70.7
HandOcc(V) [3]	25.2	53.4	24.5	54.4	25.4	72.6
Our IFVONet	**24.8**	**54.6**	**24.1**	**55.6**	**26.1**	**73.9**

C. Evaluation Metrics

J-PE/V-PE stands for joints/vertices position errors in millimetres (mm) that is achieved by calculating the average Euclidean distance between predicted joint/vertices and ground-truth 3D hand joint/vertices coordinates. Furthermore, the joints/vertices position errors are also measured after Procrustes Alignment to reflect the reconstruction quality of the hand shape.

J-AUC/V-AUC represents the area under the curve (AUC) of the percentage of correct key points (PCK) for joints and vertices, respectively. These metrics provide a comprehensive evaluation of the IFVONet's accuracy in predicting the correct positions of key points and vertices on the hand mesh. Specifically, PCK measures the fraction of key points that fall within a certain distance threshold from the ground truth, offering a robust indication of the IFVONet's precision in localizing joints and vertices.

The F-score calculates the harmonic mean of recall and precision between the predicted and ground-truth hand-mesh vertices. This metric balances the trade-off between precision (the proportion of predicted vertices that are correctly located) and recall (the proportion of ground-truth vertices that are correctly predicted). By combining these two aspects, the F-score provides a single, unified measure of the IFVONet's performance in accurately reconstructing the hand mesh. In our experiments, thresholds of 5mm and 15mm are adopted.

D. Superior Performance in Reconstructing 3D Hand

We qualitatively and quantitatively evaluate the effectiveness of IFVONet in reconstructing 3D hand, comparing it against seven baseline methods across three datasets: HO3D-v2, HO3D-v3, and 100DOH. For the baseline HandOcc [3], we extend it to a video version with two adjacent image inputs, naming it HandOcc(V).

Table III
ALIGNED RESULTS ON HO3D-V3 DATASET. **BEST** AND SECOND-BEST SCORES.

Approaches	J-PE ↓	J-AUC ↑	V-PE ↓	V-AUC ↑	F@5 ↑	F@15 ↑
ArtiBoost [36]	10.8	78.5	10.4	79.2	50.7	94.6
Keypoint Trans [37]	10.9	78.5	-	-	-	-
H2ONet [12]	9.8	80.3	9.8	80.5	51.6	95.3
Our IFVONet	**9.6**	**80.9**	**9.5**	**81.0**	**52.6**	**95.4**

①**IFVONet Consistently Outperform Baselines.** Quantitively, we report J-PE, J-AUC, V-PE, V-AUC, F@5 and F@15 on HO3D-v2 values on HOD3-v2 and HOD-v3 datasets. The results in Table I and Table II show that our proposed IFVONet achieves superior performance across almost all metrics for both before and after Procrustes Alignment (PA) cases. Notably, IFVONet consistently outperforms the baselines by a significant margin in all metrics in Table II, demonstrating the superiority of our proposed method in 3D hand reconstruction. To further validate the effectiveness of IFVONet, we conducted experiments on a recent larger hand reconstruction dataset, HO3D-v3. The results in Table III indicate that IFVONet consistently achieves better performance than the baselines, further demonstrating the effectiveness of our proposed IFVONet in 3D hand reconstruction. Qualitively, we visualize the reconstructed 3D hand based on HO3D-v2 and 100DOH datasets. The results in Fig.6 demonstrate that IFVONet generates more accurate and natural hand mesh results than baselines, even within the wild dataset(100DOH). For instance, in the first image, IFVONet produces a more natural-looking thumb and complete fingers than both HandOcc(V) and H2ONet. Such improvements are evident in the reconstructed 3D hands, particularly in the areas highlighted by red circles in Fig. 6.

②**IFVONet Effectively Captures Temporal Coherence.** To validate that IFVONet can effectively capture the temporal coherence of 3D hands in adjacent images, we present visualizations of the reconstructed 3D hands in Fig. 6. The results demonstrate that the 3D hands reconstructed by IFVONet exhibit significantly greater temporal coherence across adjacent images compared to the baselines, also indicating the effectiveness of the architectural design of the proposed IFVONet.

E. Ablation Studies

We conduct ablation experiments on the HO3D-v2 dataset using J-PE, M-PE and F-scores as the evaluation metric. Specifically, the effectiveness of each component (PDV, GOR, VOF) is investigated in our model architecture. The results are shown in Table IV, where the first row demonstrates the performance of a basic model that does not consider the utilisation of inter-frame variation based on a standard Transformer with Softmax activation function instead of Sigmoid function. The remaining rows in Table IV show the results of adding sub-components to

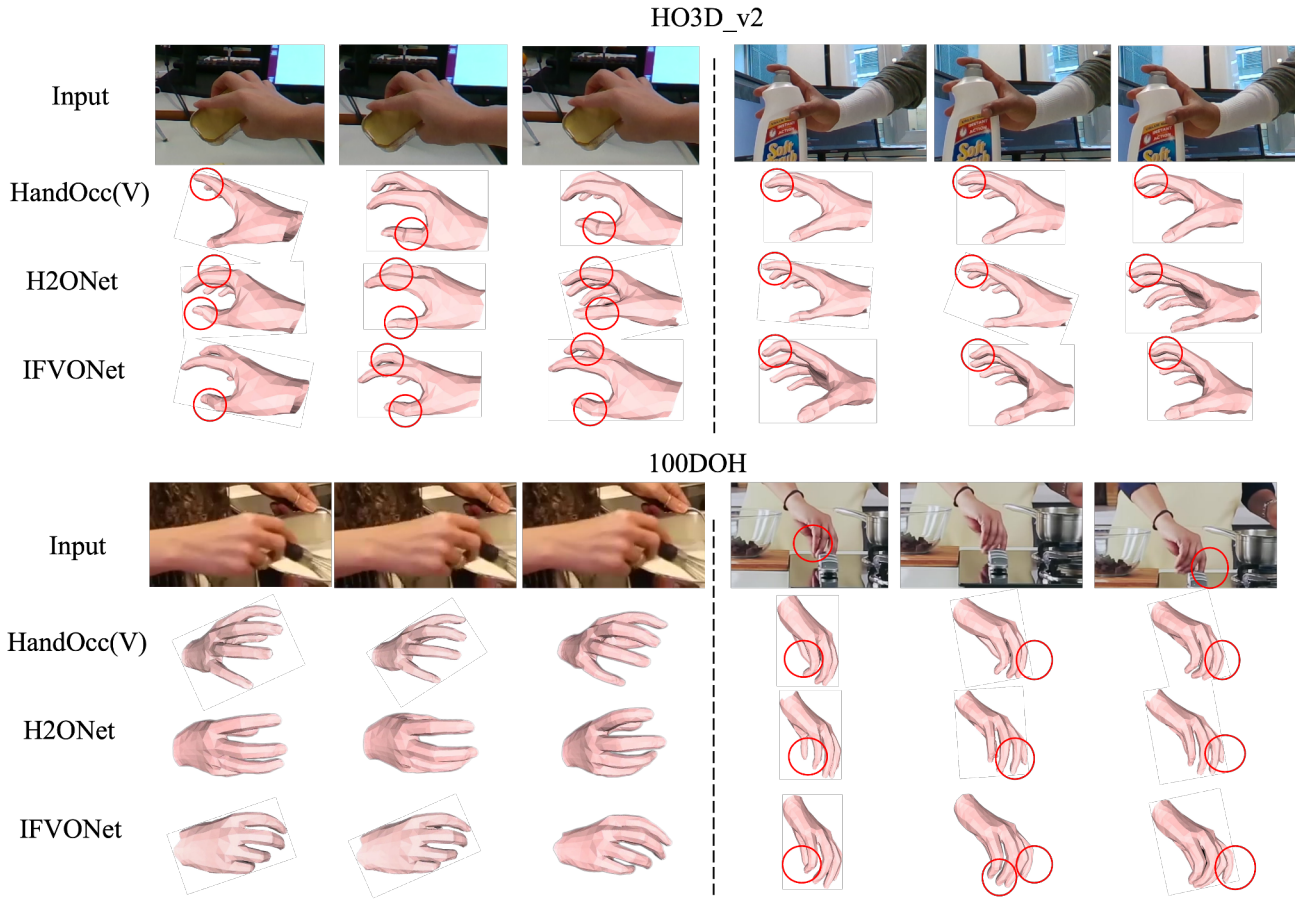

Figure 6. Qualitative result on HO3D-v2 and 100DOH dataset compared with the state-of-the-art methods.

the basic model. Our proposed `IFVONet` outperforms the basic model, indicating that each component leads to more accurate 3D hand mesh reconstruction and pose estimation.

Table IV
EFFECTIVENESS OF EACH COMPONENT ON HO3D-V2 (AFTER PA). TF: STANDARD TRANSFORMER ARCHITECTURE WITH SOFTMAX FUNCTION.

Models	J-PE ↓	M-PE ↓	F@5 ↑	F@15↑
TF	9.7	9.7	52.5	95
GOR	9.5	9.4	52.5	95.7
GOR + PDV	9.3	9.3	54.4	95.8
GOR + PDV + VOF	**9.0**	**9.0**	**55.4**	**96.2**

Ablation on GOR Table V shows the difference between FIT [3] and demonstrates that our proposed combination of Sigmoid and Batch Normalisation based attention on the F_{occ} and F_{hand} regions achieves the best performance on J-PE, M-PE and F-scores. Compared with the standard Transformer architecture (first row Table V), we can see that using the Softmax attention function declines the model performance. We also report the results of the Softmax and Batch normalisation attention-based module. This combination produces better results when compared with FIT by applying softmax attention on F_{occ} and Batch Normalisation attention on F_{hand} separately. Furthermore, to explore more information from F_{occ}, we adopt sigmoid attention instead of softmax attention, demonstrating the best performance on all the performance metrics, especially on M-PE. Therefore, we demonstrate that the Sigmoid function and Batch normalization attention can boost the model performance, indicating the effectiveness of our GOR design.

Table V
COMPARISON OF MODELS WITH VARIOUS GOR ARCHITECTURES ON HO3D-V2. TF: STANDARD TRANSFORMER ARCHITECTURE WITH INTEGRATED PDV AND VOF.

GOR architectures	J-PE ↓	M-PE ↓	F@5 ↑	F@15↑
Softmax attn. (F_{occ}, F_{hand}) (TF)	9.5	9.5	52.6	95.7
Softmax attn. (F_{occ}) + Sigmoid attn. (F_{hand}) (FIT)	9.2	9.3	54.2	95.9
Softmax attn. (F_{occ}) + BN attn. (F_{hand})	9.1	9.1	55.2	96.2
Sigmoid attn. (F_{occ}) + BN attn. (F_{hand}) (Ours)	**9.0**	**9.0**	**55.4**	**96.2**

Ablation on PDV Table VI shows that the inter-frame

variation consideration in the architecture boosts the performance between two adjacent frames. We first compare the performance with and without the inter-frame variation for two adjacent frames. Compared the results when no inter-frame variation is considered (first row Table VI), the inter-frame variation is distinguished, leading to a more coherent hand mesh result by retaining the unchanged regions. Furthermore, we compare the inter-frame variation across two frames and three frames. The inter-frame variation between three frames is implemented by interacting pixels from the nearest adjacent frames for the current frame and distinguishing the variation between the current frame and the last two frames. Performance evaluations show that the inter-frame variation between three adjacent frames results in worse performance due to overwhelming information, when compared with two adjacent inputs.

Ablation on VOF Table VII shows the effectiveness of dynamic fusion between inter-frame variation and global occlusion information. When compared with the dynamic fusion mechanism, the usage of direct addition and multiplication between inter-frame variation and global occlusion features results in worse performance, since the model cannot leverage the importance of inter-frame variations and occlusion features. However, adopting dynamic fusion and possibility scores calculation improves the performance, thereby revealing the efficiencies of adaptively combining the global occlusion and inter-frame variation information.

Table VI
COMPARISON OF MODELS WITH VARIOUS PDV ARCHITECTURES ON HO3D-V2.

PDV architectures	input frames	J-PE ↓	M-PE ↓	F@5 ↑	F@15↑
Without inter-frame variation	2	9.4	9.4	53.6	95.8
With inter-frame variation (Ours)	**2**	**9.0**	**9.0**	**55.4**	**96.2**
With inter-frame variation	3	9.6	9.6	52.4	95.4

Table VII
COMPARISON OF MODELS WITH VARIOUS VOF ARCHITECTURES ON HO3D-V2.

VOF architectures	J-PE ↓	M-PE ↓	F@5 ↑	F@15↑
Addition	9.6	9.5	52.6	96.4
Multiplication	9.4	9.4	53.3	95.9
Dynamic Fusion (Ours)	**9.0**	**9.0**	**55.4**	**96.2**

V. CONCLUSION

In this paper, we proposed a novel hand mesh framework named IFVONet, which parses hand comprehensive representation from inter-frame variation and occlusion information. Furthermore, to better leverage the balance between inter-frame variation and global occlusion information, our proposed IFVONet utilised a dynamic feature fusion mechanism that makes hand features more robust to occlusion and maintains the temporal hand feature in the sequence data. Experimental results on the 3D hand mesh benchmark achieved the latest state-of-the-art 3D hand reconstruction performance.

In future work, we intend to explore inter-frame variation from long historical frames for 3D hand reconstruction and extend the proposed methodology to other domains, such as human pose estimation. Specifically, we intend to further optimize the effectiveness of the proposed module in challenging occlusion scenarios, such as darkness environments or multi-hand occlusion scenarios.

ACKNOWLEDGMENT

This work has been partially supported by the SLAIDER project funded by the UK Research and Innovation, the UK Engineering and Physical Sciences Research Council Grant EP/Y018281/1 and the Graduate Teaching Assistantship of the University of Leicester.

REFERENCES

[1] S. Han, B. Liu, R. Cabezas, C. D. Twigg, P. Zhang, J. Petkau, T.-H. Yu, C.-J. Tai, M. Akbay, Z. Wang *et al.*, "Megatrack: monochrome egocentric articulated hand-tracking for virtual reality," *ACM Transactions on Graphics (ToG)*, vol. 39, no. 4, pp. 87–1, 2020.

[2] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 873–10 883.

[3] J. Park, Y. Oh, G. Moon, H. Choi, and K. M. Lee, "Handoccnet: Occlusion-robust 3d hand mesh estimation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1496–1505.

[4] Y. Chen, Z. Tu, D. Kang, L. Bao, Y. Zhang, X. Zhe, R. Chen, and J. Yuan, "Model-based 3d hand reconstruction via self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 451–10 460.

[5] 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.

[6] Y. Oh, J. Park, J. Kim, G. Moon, and K. M. Lee, "Recovering 3d hand mesh sequence from a single blurry image: A new dataset and temporal unfolding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 554–563.

[7] L. Zhou, Y. Chen, Y. Gao, J. Wang, and H. Lu, "Occlusion-aware siamese network for human pose estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 396–412.

[8] Z. Jiang, H. Rahmani, S. Black, and B. M. Williams, "A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 758–767.

[9] Z. Tu, Z. Huang, Y. Chen, D. Kang, L. Bao, B. Yang, and J. Yuan, "Consistent 3d hand reconstruction in video via self-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, p. 9469–9485, Aug. 2023. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2023.3247907

[10] J. Yang, H. J. Chang, S. Lee, and N. Kwak, "Seqhand: Rgb-sequence-based 3d hand pose and shape estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020, pp. 122–139.

[11] S. Liu, H. Jiang, J. Xu, S. Liu, and X. Wang, "Semi-supervised 3d hand-object poses estimation with interactions in time," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 687–14 697.

[12] H. Xu, T. Wang, X. Tang, and C.-W. Fu, "H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 048–17 058.

[13] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv preprint arXiv:2201.02610*, 2022.

[14] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3d with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9826–9836.

[15] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7122–7131.

[16] C. Jiang, Y. Xiao, C. Wu, M. Zhang, J. Zheng, Z. Cao, and J. T. Zhou, "A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose estimation from a single rgb image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8846–8855.

[17] Y. Cai, L. Ge, J. Cai, and J. Yuan, "Weakly-supervised 3d hand pose estimation from monocular rgb images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 666–682.

[18] D. U. Kim, K. I. Kim, and S. Baek, "End-to-end detection and pose estimation of two interacting hands," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 189–11 198.

[19] S. Khamis, J. Taylor, J. Shotton, C. Keskin, S. Izadi, and A. Fitzgibbon, "Learning an efficient model of hand shape variation from depth images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2540–2548.

[20] J. Malik, I. Abdelaziz, A. Elhayek, S. Shimada, S. A. Ali, V. Golyanik, C. Theobalt, and D. Stricker, "Handvoxnet: Deep voxel-based network for 3d hand shape and pose estimation from a single depth map," 2020.

[21] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid, "Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 571–580.

[22] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1831–1840.

[23] H. Zixuan, S. Leilei, S. Qiang, L. Lu, H. Xiuliang, L. Bing, and Y. Lu, "A collaborative filtering algorithm based on deep denoising auto-encoder and attention mechanism," *Computing and Informatics*, 2024.

[24] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 713–728.

[25] Y. Cheng, B. Yang, B. Wang, and R. T. Tan, "3d human pose estimation using spatio-temporal networks with explicit occlusion training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 631–10 638.

[26] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "How robust is 3d human pose estimation to occlusion?" *arXiv preprint arXiv:1808.09316*, 2018.

[27] Y. Cheng, B. Yang, B. Wang, W. Yan, and R. T. Tan, "Occlusion-aware networks for 3d human pose estimation in video," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 723–732.

[28] A. Ziani, Z. Fan, M. Kocabas, S. Christen, and O. Hilliges, "Tempclr: Reconstructing hands via time-coherent contrastive learning," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 627–636.

[29] Y. Liu, Z. Shao, Y. Teng, and N. Hoffmann, "Nam: Normalization-based attention module," *arXiv preprint arXiv:2111.12419*, 2021.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3196–3206.

[33] S. Hampali, S. D. Sarkar, and V. Lepetit, "Ho-3d_v3: Improving the accuracy of hand-object annotations of the ho-3d dataset," *arXiv preprint arXiv:2107.00887*, 2021.

[34] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," 2019.

[35] H. Qi, C. Zhao, M. Salzmann, and A. Mathis, "Hoisdf: Constraining 3d hand-object pose estimation with global signed distance fields," *arXiv preprint arXiv:2402.17062*, 2024.

[36] L. Yang, K. Li, X. Zhan, J. Lv, W. Xu, J. Li, and C. Lu, "Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2750–2760.

[37] 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.