HandOccNet: Occlusion-Robust 3D Hand Mesh Estimation Network

JoonKyu Park[1*] Yeonguk Oh[1*] Gyeongsik Moon[1*] Hongsuk Choi[1] Kyoung Mu Lee[1,2]
[1]Dept. of ECE & ASRI, [2]IPAI, Seoul National University, Korea
jkpark0825@snu.ac.kr, namepllet1@gmail.com, {mks0601, redarknight, kyoungmu}@snu.ac.kr

Abstract

Hands are often severely occluded by objects, which makes 3D hand mesh estimation challenging. Previous works often have disregarded information at occluded regions. However, we argue that occluded regions have strong correlations with hands so that they can provide highly beneficial information for complete 3D hand mesh estimation. Thus, in this work, we propose a novel 3D hand mesh estimation network HandOccNet, that can fully exploits the information at occluded regions as a secondary means to enhance image features and make it much richer. To this end, we design two successive Transformer-based modules, called feature injecting transformer (FIT) and self-enhancing transformer (SET). FIT injects hand information into occluded region by considering their correlation. SET refines the output of FIT by using a self-attention mechanism. By injecting the hand information to the occluded region, our HandOccNet reaches the state-of-the-art performance on 3D hand mesh benchmarks that contain challenging hand-object occlusions. The codes are available in: https://github.com/namepllet/HandOccNet.

1. Introduction

Despite promising results of 3D hand mesh estimation from a single RGB image [6, 12, 20, 26–30], making 3D hand mesh estimation method robust to occlusion is still an open challenge. One promising approach for the occlusion-robust system is using a spatial attention mechanism. Although the spatial attention mechanism has not been used for the occlusion-robust 3D hand mesh estimation, several 2D human body pose estimation methods [8, 39, 40] have utilized such attention mechanism for the occlusion-robust results. They estimate a spatial attention map and multiply it with a feature map to tell the networks where to focus. The attention map tends to have high scores on human regions and low scores on occluded regions. Therefore, it attenuates the magnitude of features at occluded regions and makes networks focus on human regions.

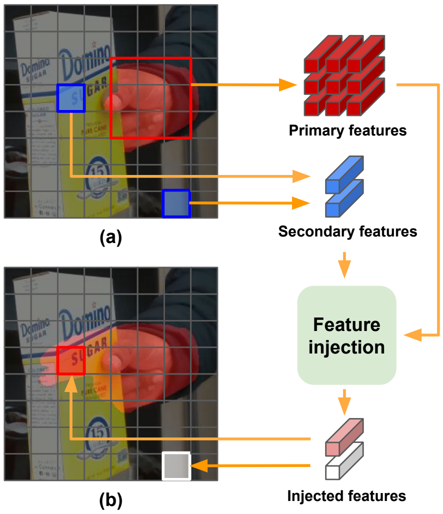


Figure 1. Example of the operation of the proposed HandOccNet. (a) The output feature map of spatial attention mechanism for the case of severe occlusion, which consists of sparse primary and secondary features . (b) Our feature injection module finds the primary features related to the secondary features, and then injects the information of primary features into the locations of secondary features.

Although the spatial attention-based methods have shown noticeable results under the occlusions, there are several limitations. First, they are mostly for the 2D human body pose estimation, which aims to localize 2D body joint coordinates. Hence, the validity of their spatial attention mechanism is not proved for the occlusion-robust 3D hand mesh estimation. In particular, as hands have quite complicated articulations and are often severely occluded by objects, the widely used spatial attention mechanism might fail to produce robust results. Unlike the methods [24, 25] using a depth map, additional depth ambiguity, which arises from 2D image-to-3D hand estimation, is another bottleneck. Second, when the occlusions are severe, activations of the spatial attention mechanism become sparse because most of the hand regions are occluded. The sparse regions contain limited information of hand; hence, relying only on

[*] Authors contributed equally.

such limited information can lead to erroneous results.

To overcome the above limitations, we propose HandOccNet, a novel framework for occlusion-robust 3D hand mesh estimation. The main component of the proposed HandOccNet is a feature injection mechanism, shown in Figure 1. The conventional spatial attention mechanism disregards the information of features at the occluded regions. On the other hand, our feature injection mechanism utilizes those features as a secondary role to obtain richer representation for occlusion-robust 3D hand mesh estimation. The *primary features* and *secondary features* represent features corresponding to high attention scores and low attention scores, respectively. We leverage information of secondary features to find relevant primary features and inject the information of primary features into the locations of secondary features. In this process, we use the term *inject* to emphasize that the information of secondary features disappears and the information of primary features is injected into empty locations.

To inject not only nearby features but also distant features, we employ Transformer [35], which has an excellent ability to model correlations between features regardless of the distance between features. Here, the distance between features represents the 2D distance in the pixel space. We build two Transformer-based modules, feature injecting transformer (FIT) and self-enhancing transformer (SET). The FIT injects the information of primary features into the regions of the secondary features and outputs a single feature map by utilizing secondary features as queries and primary features as key-value pairs. The SET utilizes a standard self-attention mechanism to refine the output of the FIT.

Our FIT has two distinctive points compared to the standard Transformer [35] for the feature injection. First, our FIT computes a correlation map between queries and keys through two types of attention modules, sigmoid-based as well as softmax-based ones, while the standard Transformer uses only softmax-based one. The softmax-based attention module normalizes the multiplications of each query and all elements of the keys using softmax function. As softmax considers all elements for the normalization, an undesirable high correlation score can be made when absolute values of all the multiplications are very low but some multiplications are relatively large compared to others. To prevent such undesirable high correlation scores, we build an additional sigmoid-based attention module. As the sigmoid activation function does not consider other elements for the normalization, it can avoid the undesired high correlations. We obtain the final correlation map by multiplying correlation maps from the softmax-based module and sigmoid-based module. Second, we remove a residual connection between input queries and output of the attention module, while the standard Transformer uses such residual connection. In other words, the FIT uses queries only when computing correlations between queries and keys and the output feature of FIT does not contain the information of the queries. This is because we intend secondary features (queries) to be replaced with primary features (values).

We demonstrate the effectiveness of our HandOccNet, through extensive experiments on recently published hand-object interaction datasets, such as HO-3D [13] and FPHA [11]. These datasets contain various and challenging occlusions in hand regions which reflects realistic occlusions that occur when hands manipulate objects in our daily life. The experimental results show that our HandOccNet achieves significantly better 3D hand mesh estimation accuracy compared to previous state-of-the-art 3D hand mesh estimation methods.

To summarize, we make the following contributions:

- We propose a HandOccNet, a novel framework for occlusion-robust 3D hand mesh estimation from a single RGB image. The proposed HandOccNet utilizes feature injection mechanism that makes feature map robust to occlusion by properly injecting the hand information into the occluded regions.
- For the feature injection and refinement, we propose two Transformer-based modules, FIT and SET. The FIT performs the injection mechanism under the guidance of correlations between primary features and secondary features, which represent features of hand regions and occluded regions, respectively. The SET refines the output feature map of the FIT using a self-attention mechanism.
- We show our framework significantly outperforms state-of-the-art 3D hand mesh estimation methods on hand-object interaction datasets that contain severe hand occlusions.

2. Related works

Occlusion-robust human pose estimation. There are three main approaches for occlusion-robust human pose estimation. The first one adopts occlusion-aware data augmentation, the second one leverages temporal information, and the last one utilizes a spatial attention mechanism.

[3, 18, 34] applied occlusion-aware data augmentation in the training time. Sarandi *et al*. [34] covered partial region of the image with black solid shapes or object segments from Pascal VOC 2012 [10] to mimic the occlusions. Ke *et al*. [18] copy background patch of the input image and paste it to human keypoint region. [3,7] proposed a two stage approach for the 3D pose estimation. They estimate 2D features for given frames and estimate the 3D pose from the 2D information. Cheng *et al*. [3] utilized sequential 2D features (2D joint heatmaps) to estimate the consecutive 3D pose. In training time, [3] randomly mask part of estimated 2D joint heatmaps by setting their values to zero in order to simulate occlusions. The limitation of their augmentations is that the occlusions are synthetic.

[4, 5] utilized temporal information to compensate for the missing information due to the occlusion. Choi *et al.* [5] and Cheng *et al.* [4] leveraged temporal information for temporally consistent mesh recovery and the occlusion-robust 3D human pose estimation from a video, respectively. [4] first estimated an incomplete 2D pose sequence, which means several joints are labeled as occluded and their coordinates are set to zero, from the input video. Then they lifted the incomplete 2D pose sequence to complete 3D pose sequence through successive 2D and 3D temporal convolutional networks.

[8, 39, 40] utilized spatial attention mechanism for the occlusion-robust system. Chu *et al.* [8] proposed a multi-context, multi-resolution and hierarchical spatial attention scheme for the 2D human pose estimation. They reweighted the feature map through their spatial attention scheme and boost 2D human pose estimation performance. Zhu *et al.* [40] first estimated a spatial attention map and multiply it by the feature map to filter out the features of occluded regions. Then they used inter-feature correlations through a shared structural matrix in order to recover missing features. Zhou *et al.* [39] also estimated the spatial attention map to filter out features of occluded regions. Then they recovered features through dilated convolutions.

Ours is related to the spatial attention mechanism; however, there are two main differences compared to the above spatial attention mechanism-based methods. First, the above methods are mostly designed for 2D human body pose estimation, which is less ambiguous than 3D hand mesh estimation that suffers from depth ambiguity and severe occlusions by objects. Second, we propose a new feature injection mechanism, which produces highly rich features even when hands are severely occluded.

3D hand mesh estimation under hand-object interaction scenarios. After hand object interaction benchmark datasets, such as HO-3D [13] and FPHA [11], had been released, several studies [13–15, 23] have been conducted on these datasets. Hasson *et al.* [15] proposed novel losses to reflect physical constraints for interacting hand and object. Hampali *et al.* [13] detected 2D joint locations and fitted a hand model (*i.e.*, MANO [33]) parameters by minimizing their loss function. Hasson *et al.* [14] leveraged a photometric consistency between neighboring frames. They estimated mesh for hand and object and rendered it to regress warping flow. Then they applied a pixel-level loss to enforce photometric consistency between a reference frame and warped frame by the regressed flow. Liu *et al.* [23] proposed a contextual reasoning module that enhances object representations by utilizing interaction between the hand and object. Most of the above methods focused on modeling interactions between hands and objects. On the other hand, we firstly introduce a novel feature injection mechanism for the occlusion-robust 3D hand mesh estimation.

Transformers. Transformers [35] showed superior results on natural language processing (NLP). Recently, vision researchers have applied Transformers to various applications, such as object detection [1], image classification [9] and human texture estimation [36]. In the field of 3D human pose and shape estimation, [17, 21, 37, 38] designed Transformer-based modules. Huang *et al.* [17] proposed Transformer-based networks which estimate 3D hand pose from 3D hand point cloud. Lin *et al.* [21] adopted a Transformer to model global vertex-to-vertex interactions and reconstructed 3D human mesh from a single RGB image. Zheng *et al.* [38] employed spatial and temporal Transformers for 3D human pose estimation in videos. Yang *et al.* [37] utilized a Transformer to capture image-specific spatial dependencies between keypoints and estimated 2D human pose. Recently, Liu *et al.* [23] proposed a Transformer-based contextual reasoning module. When an object is interacting with a hand in the input image, the contextual reasoning module enhances object regions' features by utilizing hand regions' features. The enhanced object feature is used only for the 6D object pose estimation, not for the 3D hand mesh estimation. Liu *et al.* [23] is the most relevant work with ours; however, their contextual reasoning module is used only for the 6D object pose estimation. On the other hand, our injected features are used for the 3D hand mesh estimation.

3. HandOccNet

In Figure 2, we provide an overall pipeline of our HandOccNet for 3D hand mesh estimation. Our HandOccNet consists of backbone, FIT, SET and regressor.

3.1. Backbone

The backbone extracts feature $\mathbf{F}$ and necessity map $\mathbf{M}$ from a hand images $\mathbf{I} \in \mathbb{R}^{512\times512\times3}$. We first feed the hand image $\mathbf{I}$ to ResNet50 [16]-based FPN [22] and resize the output of FPN, which produces a feature map $\mathbf{F} \in \mathbb{R}^{32\times32\times256}$. Then, we obtain a necessity map $\mathbf{M}$ from the feature map $\mathbf{F}$. We build three consecutive convolution layers, followed by the sigmoid function to estimate the necessity map $\mathbf{M}$ without supervision so that feature importance could be predicted from learning. The necessity map $\mathbf{M}$ represents scores according to spatially varying importance, which is caused by redundant information (*i.e.*objects and background) in feature $\mathbf{F}$. Using the necessity map $\mathbf{M}$, we separate the feature map $\mathbf{F}$ into primary feature $\mathbf{F}_\text{P}$ and secondary feature $\mathbf{F}_\text{S}$ with sum-to-one constraints:

$$\mathbf{F}_\text{P} = \mathbf{F} \otimes \mathbf{M},$$

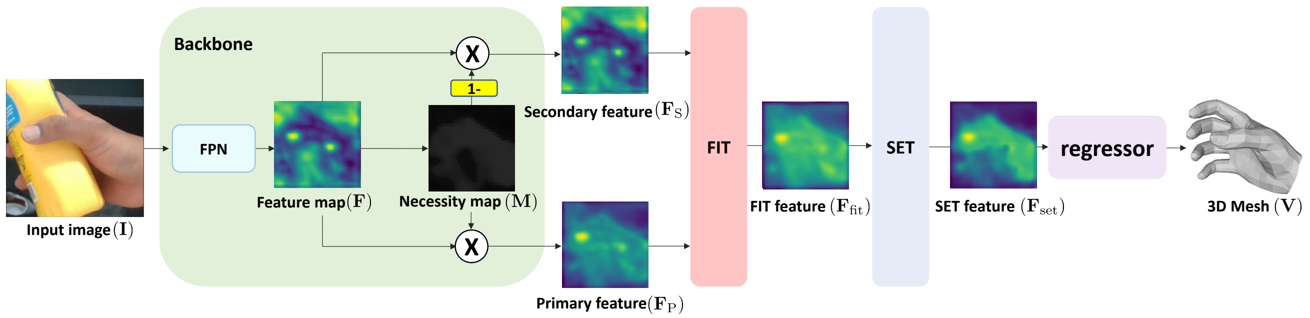

Backbone
FPN
Input image $(\mathbf{I})$
Feature map $(\mathbf{F})$
Necessity map $(\mathbf{M})$
1-
Secondary feature $(\mathbf{F}_\mathrm{S})$
Primary feature $(\mathbf{F}_\mathrm{P})$
FIT
FIT feature $(\mathbf{F}_\mathrm{fit})$
SET
SET feature $(\mathbf{F}_\mathrm{set})$
regressor
3D Mesh $(\mathbf{V})$

Figure 2. The overall architecure of HandOccNet, which consists of backbone, FIT, SET, and regressor. Our HandOccNet extracts primary feature $\mathbf{F}_\mathrm{P}$ and secondary feature $\mathbf{F}_\mathrm{S}$ using a spatial attention mechanism. Then, it uses FIT to inject the information of the primary feature $\mathbf{F}_\mathrm{P}$ into the secondary feature $\mathbf{F}_\mathrm{S}$. SET refines the output of FIT via self-attention machnism. Finally, regressor produces MANO parameters. The final 3D hand mesh is obtained by forwarding the MANO parameters to MANO layer. The cross mark in a circle represents an element-wise multiplication.

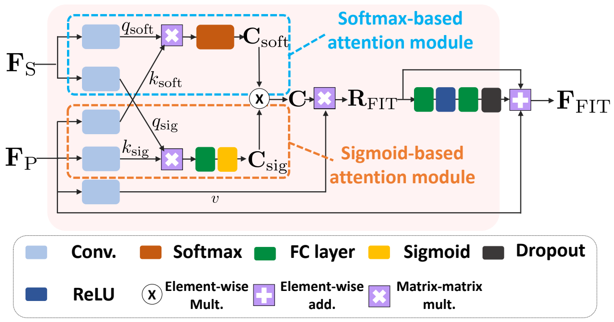


Figure 3. The overall pipeline of FIT. FIT injects the primary feature $\mathbf{F}_\mathrm{P}$ into the secondary feature $\mathbf{F}_\mathrm{S}$ using softmax-based attention module and sigmoid-based attention module.

$$\mathbf{F}_\mathrm{S} = \mathbf{F} \otimes (1 - \mathbf{M}).$$

$\otimes$ denotes element-wise multiplication. Note that $\mathbf{F}_\mathrm{P}$ contains hand regions' information, which is primarily used for hand mesh estimation and $\mathbf{F}_\mathrm{S}$ contains occluded regions' information which is not directly used for hand mesh estimating. $\mathbf{F}_\mathrm{P}$ and $\mathbf{F}_\mathrm{S}$ are utilized as query, key, and value for the following FIT.

3.2. Feature injecting transformer (FIT)

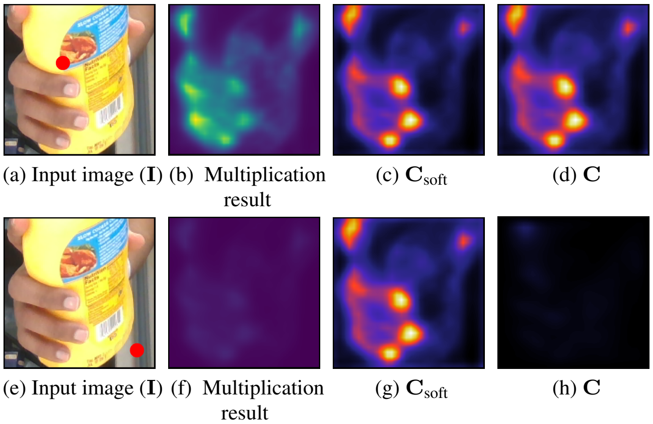
The illustration of FIT is shown in Figure 3. FIT is a Transformer-based module which takes two features, $\mathbf{F}_\mathrm{P}$ and $\mathbf{F}_\mathrm{S}$, and injects the information of $\mathbf{F}_\mathrm{P}$ into $\mathbf{F}_\mathrm{S}$ by considering their correlation. We adopt two sub-modules in the FIT called the softmax-based attention module and sigmoid-based attention module. The different role of each module is described as follows.

Softmax-based attention module. The softmax-based attention module finds the most relevant information of the primary feature $\mathbf{F}_\mathrm{P}$ from the secondary feature $\mathbf{F}_\mathrm{S}$. This can be thought as searching for the related hand information in the primary feature $\mathbf{F}_\mathrm{P}$ from the occlusion. Some object information, causing occlusion, can have strong correlation with hand information so that $\mathbf{F}_\mathrm{S}$ can tell where to inject the primary feature $\mathbf{F}_\mathrm{P}$. Therefore, while previous works utilized only $\mathbf{F}_\mathrm{P}$ and suppressed $\mathbf{F}_\mathrm{S}$ to concentrate on hand information, we use $\mathbf{F}_\mathrm{S}$ as a means of dragging and using $\mathbf{F}_\mathrm{P}$.

We extract query q_soft from $\mathbf{F}_\mathrm{S}$ and key k_soft from $\mathbf{F}_\mathrm{P}$ by two 1×1 convolution layer. Then we reshape the query and key to dimension $\mathbb{R}^{1024 \times 256}$, where 1024 represents the multiplication of width and height of $\mathbf{F}_\mathrm{P}$ and $\mathbf{F}_\mathrm{S}$. By recalling the attention mechanism of the previous Transformers [9,23,35], the softmax-based attention module generates the correlation map $\mathbf{C}_\mathrm{soft} \in \mathbb{R}^{1024 \times 1024}$ from the softmax function after the matrix multiplication of query q_soft and key k_soft:

$$\mathbf{C}_\mathrm{soft} = \mathrm{softmax}\left(\frac{q_\mathrm{soft} {k_\mathrm{soft}}^T}{\sqrt{d_{k_\mathrm{soft}}}}\right), \tag{1}$$

where $d_{k_\mathrm{soft}} = 256$ denotes the feature dimension of the key k_soft. The correlation map $\mathbf{C}_\mathrm{soft}$ indicates how much information is related between each pixel of query q_soft and key k_soft. In other words, $\mathbf{C}_\mathrm{soft}$ can be utilized to find which information of $\mathbf{F}_\mathrm{P}$ to use to fill the information of $\mathbf{F}_\mathrm{S}$. However, using only softmax for the activation is limited in handling correlation when the overall key information is not related to the specific query pixel. For example, some information (*i.e.*background) in secondary feature $\mathbf{F}_\mathrm{S}$ can be not related to the overall $\mathbf{F}_\mathrm{P}$ as in Figure 4e so that the multiplication result before the softmax function might show low values for all elements of key k_soft as shown in Figure 4f. Nevertheless, the softmax function approximates an absolutely small number, which is relatively larger than others, to a high score. Therefore, as shown in Figure 4g, undesired high correlation can occur from some relatively high elements, which are absolutely low. To use only the advantages seen in Figure 4c, which properly displays the correlation based on high multiplication result 4b, and handle the problems shown in Figure 4g, we build an additional


(a) Input image ($\mathbf{I}$) (b) Multiplication result (c) $\mathbf{C}_{\text{soft}}$ (d) $\mathbf{C}$

(e) Input image ($\mathbf{I}$) (f) Multiplication result (g) $\mathbf{C}_{\text{soft}}$ (h) $\mathbf{C}$

Figure 4. (a) and (e): red points represent example locations of query q_{soft} overlayed on the input image. (b) and (f): multiplication between the red points of query q_{soft} (shown in (a) and (e), respectively) and all elements of key k_{soft}. (c) and (g): $\mathbf{C}_{\text{soft}}$ calculated from (b) and (f), respectively, by applying a softmax function. (d) and (h) : $\mathbf{C}$ calculated from element-wise multiplication of sigmoid-based correlation map sig and $\mathbf{C}_{\text{soft}}$.

sigmoid-based attention module to filter the undesired high correlation score.

Sigmoid-based attention module. The sigmoid-based attention module filters the undesired high correlation by generating a correlation map between each query pixel and the global key information. We extract additional key-query pair, k_{sig} and q_{sig}, with same process of extracting k_{soft} and k_{soft}. Then, the module generates the correlation map $\mathbf{C}_{\text{sig}} \in \mathbb{R}^{1024\times 1}$ as follows:

$$\mathbf{C}_{\text{sig}} = \text{sigmoid}(\text{pool}(\frac{q_{\text{sig}} {k_{\text{sig}}}^T}{\sqrt{d_{k_{\text{sig}}}}})), \tag{2}$$

where pool denotes average pooling to aggregate correlation between each query q_{sig} and all elements of key k_{sig}. $d_{k_{\text{sig}}} = 256$ denotes the feature dimension of the key k_{sig}. The average pooling along the key dimension can make the correlation map $\mathbf{C}_{\text{sig}}$ robust to noisy correlations. We observed that removing the pooling in our sigmoid-based attention module makes our HandOccNet diverge during the training.

Unlike softmax function, which normalizes input element to a probability distribution considering the other elements of input, sigmoid only concentrates on normalizing a single element to a probability. Therefore, the sigmoid function does not suffer from the undesired high correlation problem of the softmax function by producing small attention scores from the small numbers of the multiplication result. We obtain our final correlation map $\mathbf{C} \in \mathbb{R}^{1024\times 1024}$ by using both correlation map from sigmoid and softmax based module, $\mathbf{C}_{\text{soft}}$ and $\mathbf{C}_{\text{sig}}$ like below:

$$\mathbf{C} = \mathbf{C}_{\text{soft}} \otimes \mathbf{C}_{\text{sig}}.$$

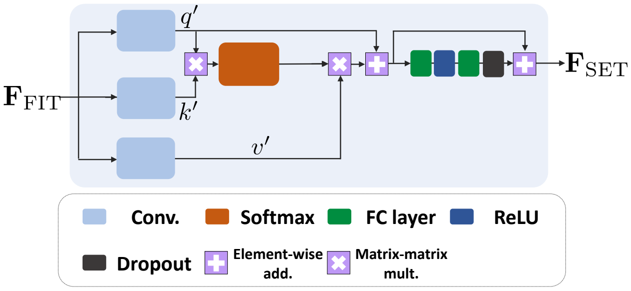


Figure 5. The overall pipeline of SET. SET refines the feature $\mathbf{F}_{\text{FIT}}$ with self-attention mechanism.

Figure 4f, 4g, and 4h show the effectiveness of the correlation map from the sigmoid-based attention module. Figure 4g shows high correlations although Figure 4f has small multiplication results, which represents that Figure 4g suffers from the undesired high correlations. By multiplying $\mathbf{C}_{\text{sig}}$ to Figure 4g, we fix the undesired high correlations, as shown in Figure 4h.

Feature injection. Using the correlation map $\mathbf{C}$, we inject the hand information to the proper occluded region. Please note that we use the word "injection" because, unlike typical Transformers [35] that use query information in output with residual connection, the query information disappears and the information of value is injected into the empty locations. We get value $v \in \mathbb{R}^{1024\times 256}$, which represents the source information indexed by the keys in Transformer, from $\mathbf{F}_{\text{P}}$ with a 1x1 convolution and flattening its spatial dimension. Then, we inject the value into the low importance region to obtain a residual feature $\mathbf{R}_{\text{FIT}} \in \mathbb{R}^{1024\times 256}$ like below:

$$\mathbf{R}_{\text{FIT}} = \mathbf{C}v. \tag{3}$$

Afterward, we feed $\mathbf{R}_{\text{FIT}}$ into a feed-forward module. The feed-forward module consists of a two-layer MLP and layer normalization with a residual connection between its input and output. We further add a residual connection between its output and the primary feature $\mathbf{F}_{\text{P}}$, which already contains essential information for hand mesh estimation. FIT's output feature $\mathbf{F}_{\text{FIT}} \in \mathbb{R}^{32\times 32\times 256}$ is obtained like below:

$$\mathbf{F}_{\text{FIT}} = \mathbf{F}_{\text{P}} + \psi(\mathbf{R}_{\text{FIT}}) + \psi(\text{MLP}(\text{LN}(\mathbf{R}_{\text{FIT}}))),$$

where ψ denotes a reshaping function that reshapes the input feature to $\mathbb{R}^{32\times 32\times 256}$. MLP and LN denote the MLP module and layer normalization layer, respectively.

3.3. Self-Enhancing transformer (SET)

The illustration of SET is shown in Figure 5. SET refines the feature $\mathbf{F}_{\text{FIT}}$ by referencing the distant information from feature $\mathbf{F}_{\text{FIT}}$ with self-attention. Different from the FIT which concentrates on injecting primary feature $\mathbf{F}_{\text{P}}$ into secondary feature $\mathbf{F}_{\text{S}}$, SET utilizes self-attention of $\mathbf{F}_{\text{FIT}}$ by extracting the query q', key k', and value v' from the same feature $\mathbf{F}_{\text{FIT}}$ with three 1x1 convolution layers. As SET performs self-attention, there is no existence of case that overall key information is not related to the query pixel because each query pixel is at least correlated to itself. Therefore, instead of using the sigmoid-based attention module which is used to filter the undesired high correlation, we only adopt the softmax-based attention module to obtain a correlation map in SET. SET follows the same pipeline of the softmax-based attention module in FIT except a residual connection between query q' and the multiplication of correlation map and value v'. The module in FIT does not have the residual connection as its goal is to "replace" the query with the value for the feature injection. On the other hand, as the goal of SET is enhancing the input feature, not the injection, we add the residual connection, following previous Transformers [35]. The output of SET is denoted by $\mathbf{F}_{\text{SET}}$. Two or more SET do not have much effect in our experiment because sufficient enhancement is already occurred in the first SET; therefore, we use one SET after the FIT.

3.4. Regressor

The regressor produces MANO pose and shape parameters, and the final 3D hand mesh is obtained by forwarding the MANO parameters to MANO layer. First, a single block of hourglass network [31] takes enhanced feature $\mathbf{F}_{\text{SET}}$ as input and outputs 2D heatmaps for each joint $\mathbf{H}$. Then, four residual blocks [16] takes a concatenation of the enhanced hand feature $\mathbf{F}_{\text{SET}}$ and the 2D heatmap $\mathbf{H}$. Finally, the output of the residual blocks are vectorized into a 2048 dimensional vector and passed to fully-connected layers, which predict MANO pose parameters $\theta \in \mathbb{R}^{48}$ and shape parameters $\beta \in \mathbb{R}^{10}$. We multiplied the joint regression matrix to a 3D mesh in rest pose and applied the forward kinematics to get the final 3D hand joints coordinates and obtained the final 3D hand mesh $\mathbf{V} \in \mathbb{R}^{778\times3}$.

To train our HandOccNet, we minimize a loss function, defined as a combination of L2 distances between the predicted and ground truths $\mathbf{H}$, θ, β, $\mathbf{V}$, and J^{3D}. J^{3D} denotes a 3D hand joint coordinates, obtained by multiplying a joint regression matrix to 3D hand mesh $\mathbf{V}$, where the matrix is defined in MANO.

4. Experiments

4.1. Implementation details

All implementations were done with PyTorch [32]. We use Adam optimizer [19] with batch size 24 for our training. On HO-3D and FPHA, each model was trained with annealing the learning rate at every 10th from the initial learning rate 10^{-4}. All other details will be available in our codes.

Architectures	Joint	Mesh	F@5	F@15
Identity	10.6	10.0	52.5	94.9
Residual blocks	10.2	9.8	51.0	95.3
FIT	9.4	9.2	54.3	96.0
SET	9.8	9.6	52.6	95.3
FIT + SET (Ours)	**9.1**	**8.8**	**56.4**	**96.3**

Table 1. Comparison of models with various architectures on HO-3D.

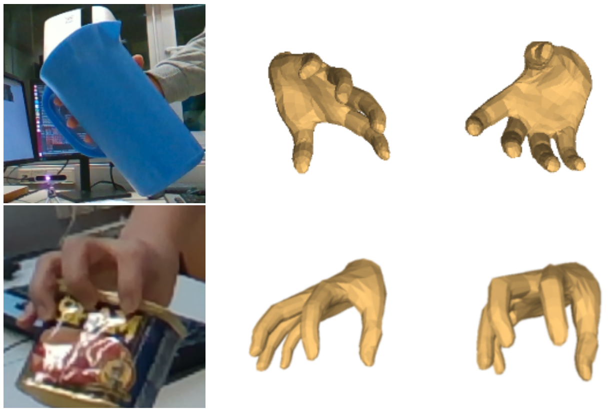

(a) Input image (**I**) (b) Wo. FIT and SET (c) W. FIT and SET (Ours)

Figure 6. Comparisons between models without and with FIT and SET on HO-3D.

4.2. Datasets and evaluation metrics

HO-3D. The HO-3D dataset [13] is a hand-object interaction dataset which contains challenging occlusions. This dataset provides RGB images with MANO-based hand joints and meshes, and camera parameters. The results on the test set can be evaluated via an online submission system.

First-Person Hand Action (FPHA). The FPHA dataset [11] contains egocentric RGB-D videos capturing a wide range of hand-object interactions. While 3D hand pose annotations are available in all frames, 6D object pose annotations are available in a small subset of the entire dataset. For the fair comparison, we follow the same train and test set split as previous works [14, 23].

Evaluation metrics. For HO3D, we report the standard metrics, such as mean joint error and mesh error in mm and F-scores, returned from the official evaluation server. For FPHA, we report the mean joint error in mm. All metrics are obtained after the procrustes alignment. Furthermore, as results before procrustes alignment are also important, we also show joint error before procrsutes alignment on the HO3D dataset in the supplementary material.

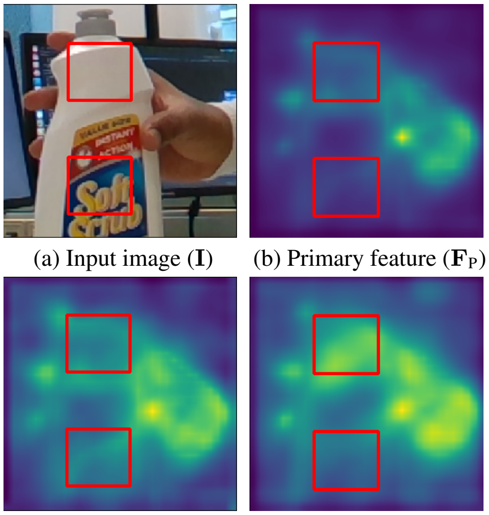

(a) Input image ($\mathbf{I}$)
(b) Primary feature ($\mathbf{F}_{\mathrm{P}}$)
(c) Output of FIT ($\mathbf{F}_{\mathrm{FIT}}$)
(d) Output of SET ($\mathbf{F}_{\mathrm{SET}}$)

Figure 7. Visualization of the feature map. Our FIT successfully injects information into the occluded region and SET makes richer information in the occluded region by self-enhancing.

4.3. Ablation study

FIT and SET. Table 1 shows that using our FIT and SET consistently improves all metrics, which demonstrates their benefits. Figure 6 further shows that FIT and SET improve the accuracy of 3D hand mesh when severe occlusions are included in the input image. For the comparison, we design four variants. All the variants have the same backbone and regressor, shown in Figure 2, and different components between the backbone and regressor. The first and second ones have a similar pipeline with that of conventional spatial attention mechanism. The first one passes the primary feature directly to the regressor, and the second one passes the primary feature to six residual blocks [16] without introducing any Transformer-based modules. They produce worse results than ours, which indicates that our newly introduced feature injection mechanism using two Transformers is highly beneficial. The third and fourth variants solely use one of FIT and SET, which produce worse results than ours. This demonstrates the efficacy of architecture of our HandOccNet using a combination of both FIT and SET.

Figure 7 shows how our FIT enhances the feature of occluded regions. Initially, red boxes in Figure 7b lack hand information due to the occlusion. Then, FIT injects hand information into the occluded region, which results in solid activation at the occluded region (red boxes), as shown in Figure 7c. Furthermore, SET enhances the information to obtain richer representation for occlusion-robust 3D hand mesh estimation, as shown in Figure 7d.

Architecture of FIT. Table 2 shows that our combination of softmax-based and sigmoid-based attention modules in FIT achieves the best results in all metrics. The sigmoid-based one filters the undesired high correlation, as shown in Figure 4. Compared to ours, using only softmax-based one like standard Transformer suffers from the undesired high correlation, which results in worse results. We also report the results of a combination of two softmax-based ones. This combination produces worse results than using a single softmax-based one, which indicates simply stacking the softmax-based ones cannot fix the undesired high correlations.

FIT architectures	Joint	Mesh	F@5	F@15
Softmax attn.	9.5	9.1	54.5	95.9
Softmax attn. + Softmax attn.	9.6	9.2	53.6	95.9
Softmax attn. + Sigmoid attn. (Ours)	**9.1**	**8.8**	**56.4**	**96.3**

Table 2. Comparison of models with various FIT architectures on HO-3D.

Settings	Joint	Mesh	F@5	F@15
Residual connection with q_{soft}	9.5	9.1	55.0	96.0
Residual connection with q_{sig}	9.7	9.3	53.3	95.7
Without residual connections (Ours)	**9.1**	**8.8**	**56.4**	**96.3**

Table 3. Comparison between models that have and do not have residual connections with query in FIT on HO-3D.

SET architectures	Joint	Mesh	F@5	F@15
Identity	9.4	9.2	54.3	96.0
Residual blocks	9.6	9.2	54.4	95.9
Single Transformer (Ours)	**9.1**	**8.8**	**56.4**	**96.3**
Two Transformers	9.2	8.9	56.2	**96.3**

Table 4. Comparison of models with various SET architecture on HO-3D.

Feature injection in FIT. Table 3 shows that removing the two residual connections achieves the best results. The first residual connection is a connection between the query of the softmax-based attention module q_{soft} and the residual feature $\mathbf{R}_{\text{FIT}}$. The second one is a connection between the query of the sigmoid-based attention module q_{sig} and the residual feature $\mathbf{R}_{\text{FIT}}$. Unlike standard Transformers, our FIT does not have the residual connection between a query and residual feature, which is a multiplication of the correlation map and value (see Eq. 3). This is because our FIT is designed to "inject" the information of value into the location of query; therefore, query is used only for the correlation map computation (see Eq. 1 and 2). The comparisons show that the residual connections are harmful for the feature injection, which results in worse performance.

Architecture of SET. Table 4 shows that designing SET as a single Transformer achieves the best results, which validates our design choice of SET. For the demonstration, we design three variants that have different SET architectures. The first one does not introduce any learnable modules in SET and just set its input feature $\mathbf{F}_{\text{FIT}}$ to the output feature

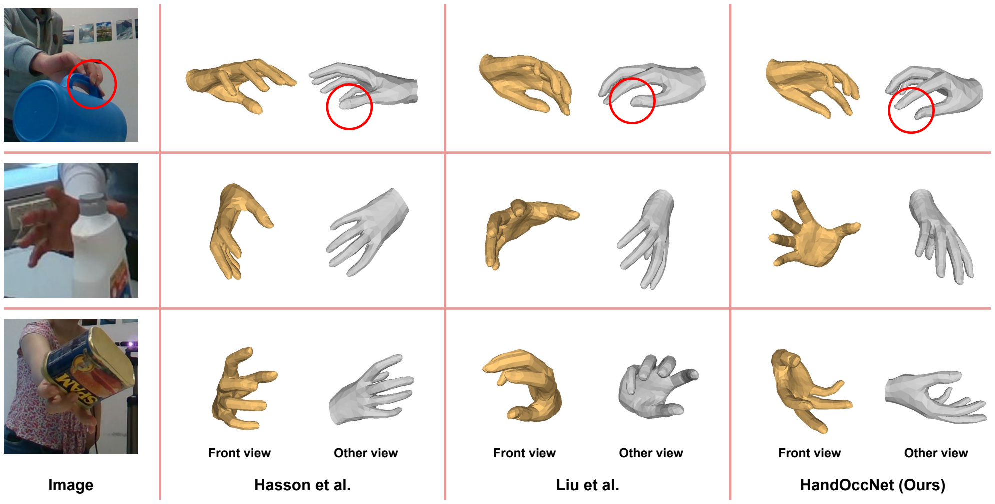


Figure 8. Qualitative comparison of the proposed HandOccNet and state-of-the-art 3D hand mesh estimation methods [14, 23] on HO-3D.

Methods	Joint	Mesh	F@5	F@15
Pose2Mesh [6]	12.5	12.7	44.1	90.9
Hasson *et al.* [14]	11.4	11.4	42.8	93.2
I2L-MeshNet [26]	11.2	13.9	40.9	93.2
Hasson *et al.* [15]	11.1	11.0	46.0	93.0
Hampali *et al.* [13]	10.7	10.6	50.6	94.2
METRO [21]	10.4	11.1	48.4	94.6
Liu *et al.* [23]	10.2	9.8	52.9	95.0
HandOccNet (Ours)	**9.1**	**8.8**	**56.4**	**96.3**

Table 5. Comparison with state-of-the-art methods on HO-3D. PA denotes Procrustes Alignment.

$\mathbf{F}_{\mathrm{SET}}$. The comparison with ours shows that the absence of learnable modules in SET produce worse results than ours, which indicates that additional feature processing is necessary. The second one uses a series of local feature extractor, which consists of three residual blocks [16]. The comparison shows that adding such local feature extractors produces worse results than ours and even worse than the first variant that does not introduce any learnable modules. This is because the newly injected features in the input feature $\mathbf{F}_{\mathrm{FIT}}$ are not locally associated. As the feature injection is performed by Transformers in FIT, distant features can be injected. Therefore, the injected features can have very different information from features of nearby pixels. Due to such locally non-associated features, the local feature extractors can have difficulty in learning local patterns, which results in worse performance. The third one uses two Transformers, which achieves slightly worse results than our single Transformer-based module. This is because a single Transformer already enhances the feature sufficiently so that the additional Transformer has a marginal effect on enhancing the input feature.

4.4. Comparisons with the state-of-the-art methods

Table 5 and 6 show that our HandOccNet achieves the best results on HO-3D and FPHA, respectively. Figure 8 shows that our HandOccNet produces much better results than state-of-the-art methods on HO-3D. As shown in the figure, our HandOccNet estimates global rotation of the hand accurately, even under the severe occlusion. Overall, our HandOccNet outperforms state-of-the-art methods on HO-3D and FPHA, which contain diverse hand-object occlusions. The results are consistent with the ablation study, which shows the proposed feature injection mechanism. Moreover, we show comparisons on larger dataset, Dex-YCB [2], to justify the efficacy of our HandOccNet in supplementary material.

Methods	3D joint error
I2L-MeshNet [26]	21.2
Hasson *et al.* [14]	18.0
Liu *et al.* [23]	16.0
Hasson *et al.* [15]	14.9
HandOccNet (Ours)	**10.8**

Table 6. Comparison with state-of-the-art methods on FPHA.

5. Conclusion

We present HandOccNet, a novel 3D hand mesh estimation framework that is robust to occlusions. Our HandOccNet utilizes a feature injection mechanism that makes feature map robust to occlusion by properly injecting the information of primary features into the location of secondary features. To this end, we design two successive Transformers: FIT and SET. Our experimental results show that our method achieves the state-of-the-art performance on 3D hand mesh benchmarks that contain severe occlusions.

6. Acknowledgment

This work was supported in part by IITP grant funded by the Korea government (MSIT) [No. 2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)].

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[2] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021. 8

[3] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3D human pose estimation using spatio-temporal networks with explicit occlusion training. In *AAAI*, 2020. 2, 3

[4] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3D human pose estimation in video. In *ICCV*, 2019. 3

[5] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *CVPR*, 2021. 3

[6] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *ECCV*, 2020. 1, 8

[7] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3D human mesh from in-the-wild crowded scenes. In *CVPR*, 2022. 2

[8] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *CVPR*, 2017. 1, 3

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 4

[10] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (VOC2012) development kit. In *PASCAL*, 2011. 2

[11] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *CVPR*, 2018. 2, 3, 6

[12] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3D hand shape and pose estimation from a single RGB image. In *CVPR*, 2019. 1

[13] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 2, 3, 6, 8

[14] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 2020. 3, 6, 8

[15] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 3, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 6, 7, 8

[17] Lin Huang, Jianchao Tan, Ji Liu, and Junsong Yuan. Hand-transformer: non-autoregressive structured modeling for 3D hand pose estimation. In *ECCV*, 2020. 3

[18] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. In *ECCV*, 2018. 2

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[20] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, 2020. 1

[21] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 3, 8

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3

[23] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3D hand-object poses estimation with interactions in time. In *CVPR*, 2021. 3, 4, 6, 8

[24] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2V-PoseNet: Voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map. In *CVPR*, 2018. 1

[25] Gyeongsik Moon, Ju Yong Chang, Yumin Suh, and Kyoung Mu Lee. Holistic planimetric prediction to local volumetric prediction for 3D human pose estimation. *arXiv preprint arXiv:1706.04758*, 2017. 1

[26] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *ECCV*, 2020. 1, 8

[27] Gyeongsik Moon and Kyoung Mu Lee. NeuralAnnot: Neural annotator for in-the-wild expressive 3D human pose and mesh training sets. *arXiv preprint arXiv:2011.11232*, 2020. 1

[28] Gyeongsik Moon and Kyoung Mu Lee. Pose2Pose: 3D positional pose-guided 3D rotational pose prediction for expressive 3D human pose and mesh estimation. *arXiv preprint arXiv:2011.11534*, 2020. 1

[29] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. DeepHandMesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *ECCV*, 2020. 1

[30] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV*, 2020. 1

[31] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 6

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[33] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. In *ACM TOG*, 2017. 3

[34] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. How robust is 3D human pose estimation to occlusion? *arXiv preprint arXiv:1808.09316*, 2018. 2

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 3, 4, 5, 6

[36] Xiangyu Xu and Chen Change Loy. 3D human texture estimation from a single image with transformers. In *ICCV*, 2021. 3

[37] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. TransPose: Keypoint localization via transformer. In *ICCV*, 2021. 3

[38] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3D human pose estimation with spatial and temporal transformers. In *ICCV*, 2021. 3

[39] Lu Zhou, Yingying Chen, Yunze Gao, Jinqiao Wang, and Hanqing Lu. Occlusion-aware siamese network for human pose estimation. In *ECCV*, 2020. 1, 3

[40] Meilu Zhu, Daming Shi, Mingjie Zheng, and Muhammad Sadiq. Robust facial landmark detection via occlusion-adaptive deep networks. In *CVPR*, 2019. 1, 3