

3D-AMTA: Occlusion-Aware Real-Time 3D Hand Pose Estimation with Auto Mask and Token-Specific Attention

Dongfang Zhao¹, Menghe Zhang¹, Yangwen Liang¹, Shuangquan Wang¹, Kee-Bong Song¹ and Donghoon Kim¹

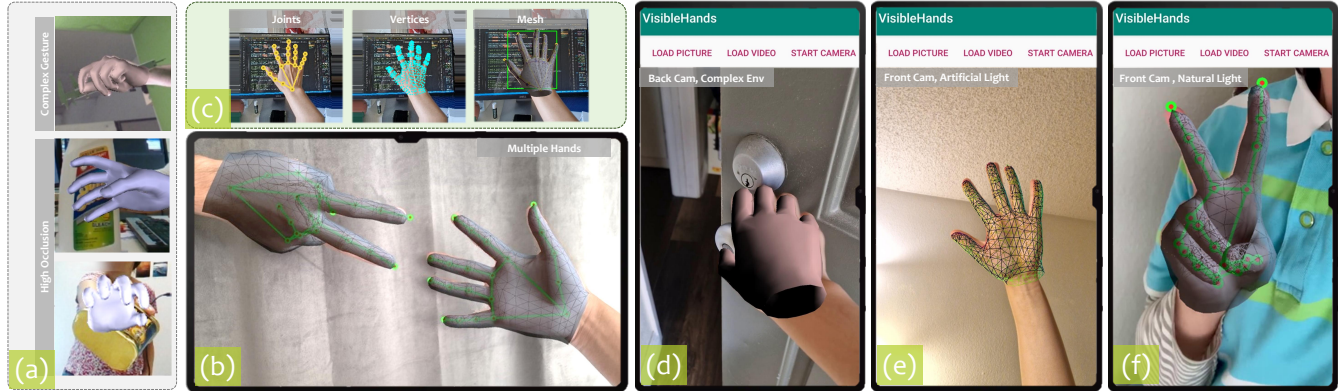


Fig. 1: Monocular 3D hand mesh estimation and reconstruction in the wild with 3D-AMTA. (a) Selected results from FreiHAND [1] and HO3D [2] showcasing complex gestures and high-occlusion scenarios. (b)–(f) 3D-AMTA integrated into our smartphone application, VisibleHands, demonstrating on-device hand reconstruction using both front and back cameras under diverse lighting conditions and user tasks.

Abstract—Understanding hand motion from a single RGB image is challenging due to occlusions and high articulation. This paper presents 3D-AMTA, a transformer-based framework with Auto Mask and Token-specific Attention for occlusion-aware 3D hand pose estimation (HPE). We propose two novel architectural enhancements: auto mask for high-occlusion scenarios, and token-specific attention for fine-grained hand articulations. These modules seamlessly integrate into transformer-based architectures that enhance real-time performance in interactive systems. To enable efficient deployment on robotic and embedded platforms, we propose 3D-AMTA-Mobile, a lightweight variant optimized for on-device processing. It achieves 267 FPS on NVIDIA RTX 2080Ti-GPU while maintaining high accuracy, making it well-suited for resource-constrained robotic applications. Extensive evaluations on FreiHAND and HO3D demonstrate that our approach consistently outperforms state-of-the-art methods in terms of accuracy, efficiency, and inference speed. These advancements contribute to robust hand perception for interactive robotics and AR-based teleoperation.

I. INTRODUCTION

Accurate 3D hand pose estimation (HPE) is essential for robotic grasping, teleoperation, and human-robot interaction [3], yet remains challenging due to occlusions and highly articulated finger motions in real-world environments. Hands frequently self-occlude, interact with objects, and exhibit

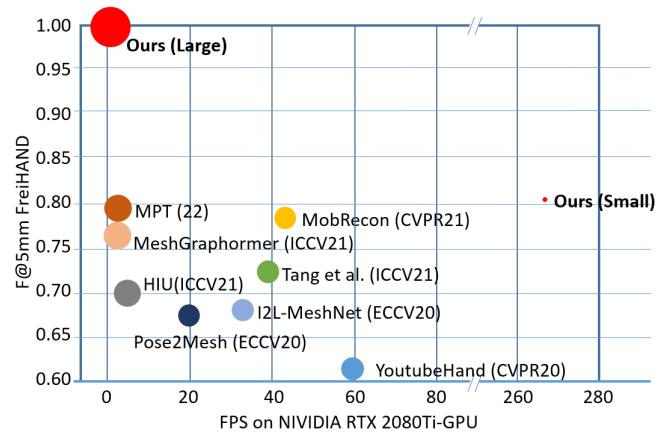


Fig. 2: 3D HPE Models Comparison: Speed vs. Accuracy. The marker size illustrates the model size.

complex deformations, which make it difficult to estimate precise 3D mesh reconstructions from a single RGB image.

Existing 3D HPEs solutions primarily rely on CNN-based architectures, while recent advancements with Graph Convolutional Neural Networks (GCNNs) and transformer modules have improved performance. However, these methods often struggle with occlusions. Standard attention aggregates features indiscriminately, so errors from occluded regions can propagate across the entire hand representation. Moreover, in real-world applications, factors such as hand scale variations

⁰¹Samsung Semiconductor Inc. SoC CMM Lab, 9868 Scranton RD, Suite 4000, San Diego, California, USA {d.zhao, menghe.z, liang.yw, shuangquan.w, keebong.s, dhoon.kim}@samsung.com

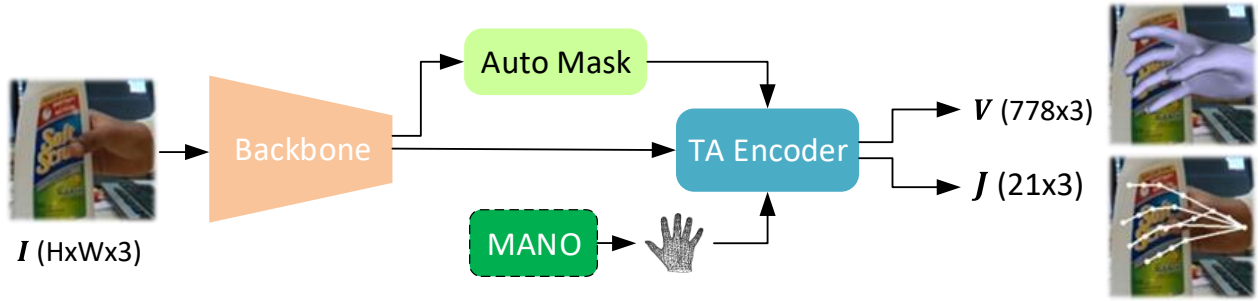


Fig. 3: **Architecture of 3D-AMTA.** The design of 3D-AMTA is based on a CNN-backbone Transformer-encoder architecture. We introduce Auto Mask module, which generates occlusion masks and applies them to tokens in the Token-specific-Attention (TA) encoder. The TA encoder integrates Transformer and GCNN modules to enhance articulated hand modeling.

and world transformations introduce ambiguities in pose estimation, commonly referred to as non-aligned hand poses. Most existing approaches neglect this issue, limiting their robustness in practical settings.

To address these challenges, we propose 3D-AMTA, a transformer-based framework for occlusion-aware 3D HPE. We introduced two key enhancements: Auto Mask Module, which dynamically suppresses unreliable features from occlusion-induced noise; and Token-specific Attention, which improves fine-grained articulation modeling by effectively capturing both global and local dependencies. These modules integrate seamlessly into transformer and graph-transformer architectures, achieving state-of-the-art performance on FreiHAND [1] and HO3D [2]. Our key contributions are:

- **Auto Mask Module.** We design a auto mask module that enhances occlusion robustness by filtering unreliable features before encoding.
- **Token-specific Attention.** We introduce a token-specific attention encoder to improve articulated hand pose modeling by capturing structural dependencies.
- **Efficient Real-Time Deployment.** We develop 3D-AMTA-Mobile, a lightweight variant optimized for embedded and mobile platforms, achieving real-time performance while maintaining high accuracy.

As shown in Figure 1, 3D-AMTA and 3D-AMTA-Mobile outperform existing methods in accuracy, with the mobile version offering superior efficiency at a significantly smaller model size. Therefore, our work enables real-time hand perception for robotics, AR, and teleoperation, bridging high-fidelity 3D modeling with real-world deployment.

II. RELATED WORKS

A. 3D HPE from a Single RGB Image

CNN-based Methods. CNNs have been the foundation of hand pose estimation due to their proficiency in capturing local features. Pioneering CNN-based works sought to extract 2D features and predict 3D hand vertices using heat-map extraction layers and keypoint regression layers [4]. Such models excel in extracting local features from RGB images,

but fall short in capturing global attention among different hand joints/vertices [5].

GCNN-based Methods. Building on the CNN framework, GCNN models incorporate graph structures to capture the relational information among joints. However, the GCN modules exponentially introduce noisy information from extended neighbor hand joints/vertices while constructing global relationships, which results in over-smoothing of the result hand meshes [6]. Pre-weight sharing GCNN is an improved mechanism that is used in various GCNN structures [7], [8]. However, the potential influence of the mechanism in transformers was not well investigated.

Transformer-based Methods. The advent of transformers introduced an alternative approach, focusing on long-range dependencies and interactions among features [9], [10], [11]. However, they encountered difficulty in capturing local features for each hand joint/vertex to construct the crucial details [5]. Despite their success, existing models [12], [4], including Graphormer [5], still face limitations in dealing with complex real-world scenarios such as occlusions and non-aligned cases. They may perform well on aligned images, where they obtain joint or vertex error metrics after aligning predicted joints or vertices with the ground truth using a predefined alignment method like Procrustes alignment [13]. However, real-world scenarios are not aligned, which means that predicted joint or vertex error metrics should be obtained by direct comparison against the ground truth. The most recent work of Huang et al. [14] addressed this issue with implicit 3D representations to locate 3D hand joints, yet the implicit method doesn't construct an explicit mesh for hand reconstruction.

B. Mask Strategy

Applying masks is popularly used especially in image inpainting problems [15] to mask the features that will act as noises or unhelpful information to the network. Meanwhile, cases with self-occlusion and object-occlusion make the HPE problem hard to solve. However, the occlusion area is not given comparing to the image inpainting problems. Therefore, we propose to estimate the occluded hand joints/vertices

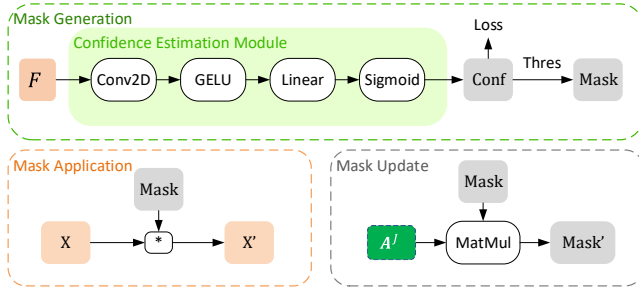


Fig. 4: **Auto Mask Module.** The mask is automatically generated from input features using the Confidence Estimation Module and applied to tokens in the TA Encoder. The graphic adjacency matrix A^J , is applied to update the mask in each encoder.

with an auto mask module with auto-generated masks to mask the features in the encoders to avoid noise features.

III. METHODS

A. Model Overview

Figure 3 illustrates the architecture of 3D-AMTA, which estimates 3D hand joints and mesh vertices from a single RGB image. To start with, a CNN-based backbone extracts image features, which are further refined by the Auto Mask module (Section III-B) to generate occlusion-aware mask features. These masked features, along with MANO hand model template, serve as inputs to the Token-specific-Attention (TA) Encoder (Section III-C). The TA Encoder is a stacked transformer-based module that extends Graphormer [5], which integrates global, local, and structural cues for more precise hand pose estimation. Additionally, if available, camera intrinsics and wrist position (or bone length) are also incorporated into token representations to improve scale-consistent 3D reconstruction even in in-the-wild images where depth ambiguity is a challenge.

B. Auto Mask Module (AM)

The backbone extracts features of hand joints and vertices for decoder prediction. However, occlusions cause incorrect features corresponding to occluders rather than targets. To address this, we propose the AM module, selectively masking tokens of occluded joints or vertices, thus improving prediction accuracy in occlusion scenarios.

Mask Generation. As shown in Figure 4, the confidence estimation module is applied to the features from the encoder to extract the confidence score of hand joint estimation. The mask is further obtained by filtering the confidence score with a threshold. Considering a confidence score to a hand vertex/joint $c_j \in \mathcal{C}$, $j \in [1, N]$, where N denotes the number of hand joints and vertices. Its ground truth $c_j^{gt} \in \mathcal{C}^{gt}$, $j \in [1, N]$ can be obtained as:

$$c_j^{gt} = \|\mathbf{p}_j^{3D} - \mathbf{p}_j^{3D,gt}\|, \forall j \in [1, N], \quad (1)$$

where $\mathbf{p}_j^{3D}$ is the j_{th} component of the estimated 3D hand pose $\mathbf{P}^{3D} \in \{\mathbf{J}, \mathbf{V}\}$ and $\mathbf{p}_j^{3D,gt}$ is the ground truth.

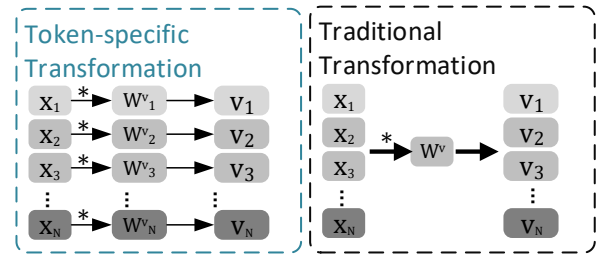


Fig. 5: **Token-Specific Transformation.** Unlike standard self-attention, where the same weight matrix applied to all tokens, our approach assigns unique weights to each token based on the features adaptively.

We set the threshold to be the average of the confidence to prevent the case when all features are masked:

$$m_j = c_j < \frac{\sum_{i=1}^{i=N} c_i}{N}, \forall j \in [1, N]. \quad (2)$$

Mask Application. We apply the mask to the encoders by multiplying the masks with the tokens in the TA encoder. In such way, the corresponding token of the hand joint/vertex will be considered as occluded and the model will estimate the position of the hand joint/vertex based on the features of the other tokens.

$$\mathbf{X}'_j = m_j * \mathbf{X}_j, \forall j \in [1, N]. \quad (3)$$

where $\mathbf{X}$ and $\mathbf{X}'$ are token features in the TA encoder (the first token feature is the feature extracted by the backbone from the input image).

Mask Update. The mask is dynamically adjusted using the graph adjacency matrix of hand joints and vertices A^J (extracted from the hand joint and vertex template by MANO):

$$\mathbf{m}' = A^J \mathbf{m} \quad (4)$$

The mask ($\mathbf{m}$) indicates whether a token is valid or invalid (includes occluded features) at the current stage. The updating strategy follows a rule that all tokens are updated after attention propagation as long as there is at least one valid adjacent token. If all adjacent tokens are invalid, they remain invalid after the current updating stage. The mask will be fully valid after several updating stages.

C. Token-specific Attention (TA)

Inspired by the pre-aggregation GCNN method [16], we extend this concept to token-level operations within the GCNN module. To further enhance feature representation, we introduce a token-specific attention, which operates across multiple stacked transformer layers, where each token is assigned a unique transformation weight at every stage.

A standard self-attention module computes the output as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V}, \quad (5)$$

Model	PJ (cm)↓	PV (cm)↓	F@5 ↑	F@15 ↑	PS (M)↓	GFLOPs↓	FPS↑
HaMuCo [17]	0.62	0.67	0.720	-	-	-	-
MobRecon [4]	0.61	0.63	0.758	0.983	32.70	4.54	76.7
Graphormer [5]	0.59	0.60	0.764	0.986	187.52	82.81	4.2
Ours	0.58	0.59	0.772	0.987	256.66	124.01	2.8
Ours (mobile)	0.78	0.79	0.336	0.837	12.65	1.30	267
†WiLoR [18]	0.55	0.51	0.825	0.993	-	-	-
†Hamba [19]	0.57	0.53	0.806	0.992	-	-	-
†MPT [12]	0.54	0.56	0.789	0.988	187.52	82.81	4.2
†MobRecon [4]	0.57	0.58	0.784	0.986	32.70	4.54	76.7
†Ours	0.19	0.17	0.993	1.000	256.66	124.01	2.8
†Ours (mobile)	0.43	0.45	0.856	0.999	12.65	1.30	267

TABLE I: Performance comparison of 3D-AMTA and 3D-AMTA-Mobile with SOTA models on FreiHAND [1] dataset. Performance reported under the aligned evaluation category with best results highlighted in bold. All benchmarks are conducted with NVIDIA RTX 2080Ti-GPU. † Indicates models trained with additional data.

Model	PJ↓	PV↓	F@5 ↑	F@15 ↑
Aligned				
Graphormer [5]	1.24	1.26	0.426	0.924
AMVUR [20]	0.87	0.83	0.593	0.964
Keypoint-T [21]	1.09	-	-	-
Ours	0.93	0.87	0.582	0.962
Non-Aligned				
NVF [14]	2.32	-	-	-
Graphormer [5]	2.80	2.60	0.233	0.700
Keypoint-T [21]	2.48	-	-	-
Ours	2.30	2.20	0.308	0.774

TABLE II: Performance comparison of 3D-AMTA with SOTA on the HO3Dv3 dataset [2]. Results are reported under both the aligned and non-aligned categories.

where Q , K , and V represent the query, key, and value matrices, respectively, and d_k denotes the hidden feature dimension.

As shown in Figure 5, traditional transformers apply the same weight matrix W^V across all tokens when computing the value matrix V from hidden features X :

$$V_t = X_t W^V \quad (6)$$

On the contrary, our token-specific attention assigns unique weight matrices W_t^V to each token:

$$V_t = X_t W_t^V, \quad (7)$$

where W_t^V is a token-specific weight matrix for the t^{th} token. This approach allows the model to adaptively transform features of each token, such as to improve the ability to capture diverse hand gesture variations more effectively.

IV. EXPERIMENTS

A. Implementation Details

3D-AMTA utilizes HRNet-w64 [22] as the backbone. The TA encoder consists of 12 layers, each with 4 multi-head self-attention (MHSA) heads. To efficiently represent hand poses, the encoder follows a progressive feature compression strategy, using hidden feature dimensions of 1024, 256, 64, and 16 across its four encoding stages. The auto-generated

mask is applied in the first two encoders, while the token-specific attention is incorporated in the third and fourth encoders.

For 3D-AMTA-Mobile, we optimize the model for real-time performance while preserving core accuracy. Instead of HRNet-w64, we adopt VAN-B0 [23] as the backbone, which maintains high feature extraction accuracy with significantly fewer parameters. To improve efficiency, we reduce the number of transformer layers in the TA encoder, starting with 2 layers and gradually increasing to 4 layers for hierarchical feature extraction. Additionally, we compress feature representations by progressively downsizing hidden feature dimensions ($512 \rightarrow 256 \rightarrow 64 \rightarrow 16$) to reduce memory usage while retaining essential pose information.

Our model was trained using the L_1 loss for 3D hand vertices and L_2 loss for 3D hand joints, 2D hand joints, and 2D hand vertices. We employed adaptive moment estimation with weight decay (AdamW) [24] as the optimizer. The 3D-AMTA was trained for 150 epochs at a learning rate of $1e^{-4}$, and for 50 epochs at a learning rate of $1e^{-5}$, with a batch size of $B = 32$ on eight NVIDIA A100 GPUs.

We performed comprehensive evaluations of 3D-AMTA and 3D-AMTA-Mobile on the FreiHAND and HO3Dv3 benchmarks to assess their performance.

B. Evaluation Criterion

We select the FreiHAND [1] and HO3Dv3 [2] datasets to evaluate our method under challenging, real-world conditions. These datasets feature diverse, in-the-wild scenarios with complex hand articulations, varying viewpoints, and heavy occlusions.

Following standard 3D hand pose estimation benchmarks, we evaluate model performance under two categories: “aligned” and “non-aligned” hand poses.

In the “aligned” category, predicted joints or vertices are adjusted using predefined alignment methods such as Procrustes alignment or translation-and-scale alignment. These techniques apply translation, scaling, and rotation to minimize the least-squares difference between predictions and ground truth, compensating for global variations. This alignment allows for a more controlled evaluation of relative hand

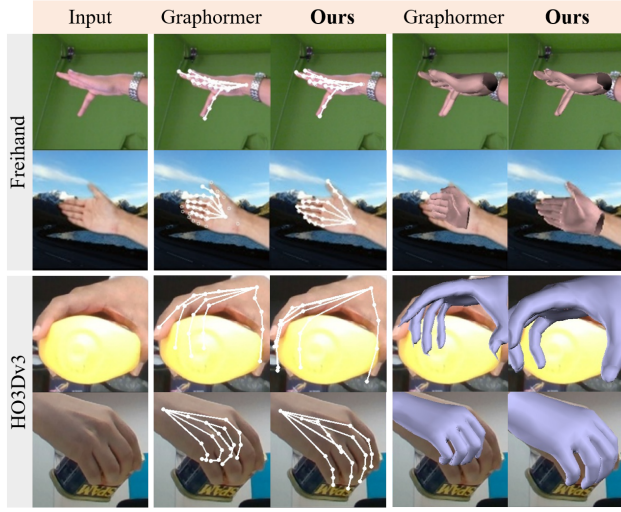


Fig. 6: Comparisons with the Graphormer [5] on FreiHand [1] and HO3Dv3 dataset. The ground truth hand joint locations are marked with white hollow circles, while the predicted locations are marked with white solid circles.

articulation accuracy, independent of hand position and scale in the scene. In the FreiHAND[1] and HO3Dv3[2] datasets, Procrustes alignment [13] is applied to align estimated hand poses to the ground truth in the aligned evaluation setup, while root-relative and scale alignment is also used in some cases.

In the “non-aligned” category, no alignment techniques are applied, meaning predictions are directly compared to the ground truth in world space. This requires the model to estimate absolute 3D hand poses, including position, scale, and orientation, making it more reflective of real-world scenarios. Performance of a model in this category demonstrates its ability to generate scale- and spatial- aware 3D hand reconstructions without relying on post-processing alignment.

Our evaluation metrics include:

PJ/PV measures the mean per joint/vertex position error by the Euclidean distance (cm) between the estimated and ground-truth coordinates;

$F@5/F@15$ measures the F-Score, the harmonic mean of recall and precision, computed between two meshes at a distance threshold of 5mm or 15mm.

$PS (M)$ and $GFLOPs$ represent the model parameter size in Million and floating-point operations in billion respectively.

FPS denotes a model’s inference speed, measured in frames per second (FPS), based on the computational capabilities of the NVIDIA RTX 2080 Ti GPU.

C. Results

Our quantitative results on the FreiHAND [1] (Table I) and HO3Dv3 [2] (Table II) datasets demonstrate the effectiveness of 3D-AMTA. On FreiHAND, 3D-AMTA achieves significantly lower PJ and PV errors while outperforming other models in $F@5/F@15$ thresholds, which highlight its improved precision in fine-grained hand mesh reconstruction.

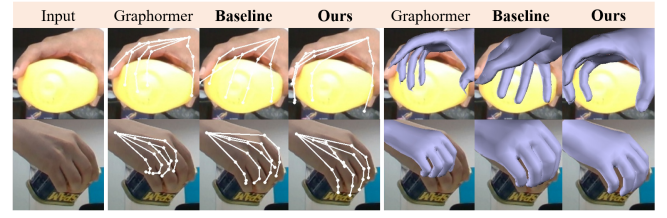


Fig. 7: Ablation study of AM module tackling high occlusion cases from HO3Dv3 dataset. Comparisons among the Graphormer [5], the Baseline (the 3D-AMTA without the AM module), and the full 3D-AMTA.

Design	E2	E3	E4	PJ↓
Baseline	x	x	x	0.590
MHSA-TA	x	x	o	0.579
	x	o	o	0.578
GCNN-TA	x	x	o	0.580
MHSA-TA	x	o	o	0.577
GCNN-TA	x	x	o	

TABLE III: Ablation study on Token-specific Attention (-TA) in MHSA and GCNN blocks on FreiHAND dataset. ‘E#’ denotes the encoder index. ‘x’ indicates that TA is not applied to the specified encoder, while ‘o’ denotes that it is applied.

In the meantime, the 3D-AMTA-Mobile outperforms the previous SOTAs while ensuring real-time mobile deployment.

On HO3Dv3, 3D-AMTA consistently reduces PJ and PV errors, further demonstrating its robustness across datasets. In challenging occlusion cases, our model exhibits notably better joint and vertex alignment compared to previous methods (Figure 6), which shows its ability to handle partially visible hands effectively.

V. ABLATION STUDY

We also conducted extensive ablation studies on the FreiHAND[1] and HO3Dv3[2] datasets to analyze the impact of our proposed design choices.

TA in Transformers and GCNNs

We compare our model against a baseline encoder that uses standard attention, where all tokens share the same transformation process without the TA. As shown in Table III, incorporating TA improves joint estimation accuracy, reducing PJ error by 0.013 cm. This demonstrates the benefits of allowing the encoder to adaptively process different token types, capturing fine-grained articulation details more effectively.

Design	E1	E2	E3	E4	PJ↓
Baseline	x	x	x	x	1.180
+camera intrinsics	x	x	x	x	0.959
+Auto Mask	o	x	x	x	0.942
	o	o	x	x	0.929
	o	o	o	x	0.946

TABLE IV: Ablation study on the AM module on HO3Dv3.

Additionally, we analyze the optimal placement of TA within the multi-stage encoding pipeline. Results indicate that applying it in the third and fourth TA encoders yields the best performance. This suggests that early-stage encoding benefits from general global attention, while later stages require more specialized feature refinement to accurately capture articulation details.

Auto Mask (AM) Module

To evaluate the effectiveness of the AM module and the inclusion of camera intrinsics, we performed experiments on the HO3Dv3 dataset. For fair comparison, we apply our TA to Graphormer [5] as the baseline. As shown in Table IV, incorporating camera intrinsics into the encoder improved hand pose estimation accuracy, particularly for hand vertices. Additionally, the AM module further reduced joint estimation error by 0.02cm. Figure 7 compares qualitative results under high-occlusion conditions among Graphormer [5], the Baseline (Graphormer w/ TA), and 3D-AMTA.

We further analyze the placement of the Auto Mask module across encoding stages. As shown in Table IV, we apply the AM progressively, from the first encoder (E1) to all four encoders (E1-E4). The results show that performance peaks when Auto Mask is applied in the first two encoders (E1-E2), with further expansion to E3-E4 leading to a decline in accuracy. This suggests that over-masking can suppress valuable tokens throughout the encoding process, preventing updates from global attention and degrading performance.

VI. CONCLUSION AND FUTURE WORK

We introduce 3D-AMTA, a transformer-based framework for occlusion-aware 3D hand pose estimation, along with 3D-AMTA-Mobile, an optimized variant for real-time deployment. Through Auto Mask module and Token-specific Attention, our method effectively handles occlusions, fine-grained articulation, and scale consistency, achieving state-of-the-art performance on FreiHAND and HO3Dv3 benchmarks.

Future work should explore temporal modeling for video-based hand tracking, improving consistency and robustness in dynamic interactions. While 3D-AMTA-Mobile demonstrates feasibility for mobile deployment, extending it to human-robot interaction and teleoperation would enhance its practical impact. Advancing time-domain optimization, on-device efficiency, and robotic perception will further strengthen its applications in real-world interactive systems.

REFERENCES

- [1] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, "FreiHAND: A dataset for markerless capture of hand pose and shape from single rgb images," in *IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2019, pp. 813–822.
- [2] 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.
- [3] R. Li, Z. Liu, and J. Tan, "A survey on 3d hand pose estimation: Cameras, methods, and datasets," *Pattern Recognition*, vol. 93, pp. 251–272, 2019.
- [4] X. Chen, Y. Liu, Y. Dong, X. Zhang, C. Ma, Y. Xiong, Y. Zhang, and X. Guo, "Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image," *arXiv preprint arXiv:2112.02753*, 2021.
- [5] K. Lin, L. Wang, and Z. Liu, "Mesh graphormer," in *IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2021, pp. 12939–12948.
- [6] D. Kong, H. Ma, and X. Xie, "Sia-gcn: A spatial information aware graph neural network with 2d convolutions for hand pose estimation," *arXiv preprint arXiv:2009.12473*, 2020.
- [7] K. Liu, R. Ding, Z. Zou, L. Wang, and W. Tang, "A comprehensive study of weight sharing in graph networks for 3d human pose estimation," in *European Conference on Computer Vision (ECCV)*, Springer, 2020, pp. 318–334.
- [8] Z. Zou and W. Tang, "Modulated graph convolutional network for 3d human pose estimation," in *IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2021, pp. 11477–11487.
- [9] D. Gao, B. Zhang, Q. Wang, X. Zhang, P. Pan, and Y. Xu, "Scat: Stride consistency with auto-regressive regressor and transformer for hand pose estimation," in *IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2021, pp. 2266–2275.
- [10] S. Hampali, S. D. Sarkar, M. Rad, and V. Lepetit, "Handformer: Keypoint transformer for monocular 3d pose estimation of hands and object in interaction," *arXiv preprint arXiv:2104.14639*, 2021.
- [11] T. H. E. Tse, F. Mueller, Z. Shen, D. Tang, T. Beeler, M. Dou, Y. Zhang, S. Petrovic, H. J. Chang, J. Taylor, et al., "Spectral graphormer: Spectral graph-based transformer for egocentric two-hand reconstruction using multi-view color images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14666–14677.
- [12] K. Lin, C.-C. Lin, L. Liang, Z. Liu, and L. Wang, "Mpt: Mesh pre-training with transformers for human pose and mesh reconstruction," *arXiv preprint arXiv:2211.13357*, 2022.
- [13] P. H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.
- [14] L. Huang, C.-C. Lin, K. Lin, L. Liang, L. Wang, J. Yuan, and Z. Liu, "Neural voting field for camera-space 3d hand pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8969–8978.
- [15] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10758–10768.
- [16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
- [17] X. Zheng, C. Wen, Z. Xue, P. Ren, and J. Wang, "Hamuco: Hand pose estimation via multiview collaborative self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20763–20773.
- [18] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou, "Wilor: End-to-end 3d hand localization and reconstruction in-the-wild," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12242–12254.
- [19] H. Dong, A. Chharia, W. Gou, F. Vicente Carrasco, and F. D. De la Torre, "Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba," *Advances in Neural Information Processing Systems*, vol. 37, pp. 2127–2160, 2024.
- [20] 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.
- [21] 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. 11090–11100.
- [22] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5693–5703.
- [23] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *arXiv preprint arXiv:2202.09741*, 2022.
- [24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.