

MaskHand: Generative Masked Modeling for Robust Hand Mesh Reconstruction in the Wild

Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Mayur Jagdishbhai Patel, Hongfei Xue, Ahmed Helmy, Srijan Das, Pu Wang
University of North Carolina at Charlotte, Charlotte, NC, USA

{msaleem2, epinyoan, mpate169, hxue2, ahelmy1, sdas24, pwang13}@charlotte.edu

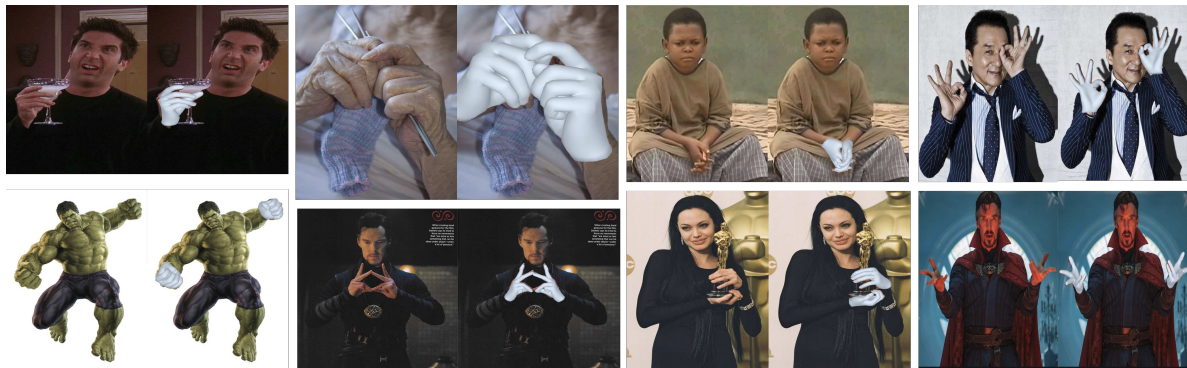


Figure 1. MaskHand: a novel generative masked model for accurate and robust 3D hand mesh recovery from single RGB images, excelling in diverse scenarios like occlusions, hand-object interactions, and varied appearances. Watch our Supplemental Video to see it in action!

Abstract

Reconstructing a 3D hand mesh from a single RGB image is challenging due to complex articulations, self-occlusions, and depth ambiguities. Traditional discriminative methods, which learn a deterministic mapping from a 2D image to a single 3D mesh, often struggle with the inherent ambiguities in 2D-to-3D mapping. To address this challenge, we propose **MaskHand**, a novel generative masked model for hand mesh recovery that synthesizes plausible 3D hand meshes by learning and sampling from the probabilistic distribution of the ambiguous 2D-to-3D mapping process. MaskHand consists of two key components: (1) a **VQ-MANO**, which encodes 3D hand articulations as discrete pose tokens in a latent space, and (2) a **Context-Guided Masked Transformer** that randomly masks out pose tokens and learns their joint distribution, conditioned on corrupted token sequence, image context, and 2D pose cues. This learned distribution facilitates confidence-guided sampling during inference, producing mesh reconstructions with low uncertainty and high precision. Extensive evaluations on benchmark and real-world datasets demonstrate that MaskHand achieves state-of-the-art accuracy, robust-

ness, and realism in 3D hand mesh reconstruction. Project website: <https://m-usamasaleem.github.io/publication/MaskHand/MaskHand.html>.

1. Introduction

Hand mesh recovery has gained significant interest in computer vision due to its broad applications in fields such as robotics, human-computer interaction [43, 57], animation, and AR/VR [12, 31]. While previous methods have explored markerless, image-based hand understanding, most depend on depth cameras [3, 24, 44, 50, 58] or multi-view images [5, 25, 54, 56]. Consequently, most of these methods are not feasible for real-world applications where only monocular RGB images are accessible. On the other hand, monocular hand mesh recovery from a single RGB image, especially without body context or explicit camera parameters, is highly challenging due to substantial variations in hand appearance in 3D space, frequent self-occlusions, and complex articulations.

Recent advances, especially in transformer-based methods, have shown significant promise in monocular hand mesh recovery (HMR) by capturing intricate hand struc-

tures and spatial relationships. For instance, METRO [15] and MeshGraphormer [40] utilize multi-layer attention mechanisms to model both vertex-vertex and vertex-joint interactions, thereby enhancing mesh fidelity. Later, HaMeR [48] illustrated the scaling benefits of large vision transformers and extensive datasets for HMR, achieving improved reconstruction accuracy. However, these methods are inherently discriminative, producing deterministic outputs for each image. Consequently, they face limitations in complex, real-world scenes where ambiguities arise due to occlusions, hand-object interactions, and challenging viewpoints are prevalent.

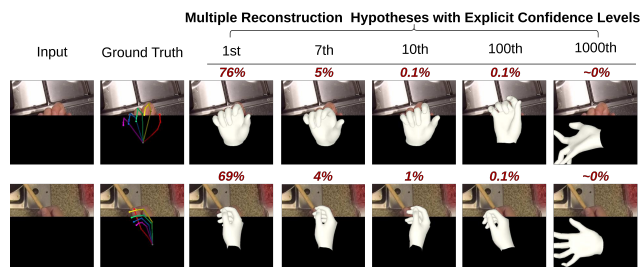


Figure 2. MaskHand is a quantifiable probabilistic HMR method that can learn 2D-to-3D mapping distributions and explicitly estimate confidence levels or prediction probabilities for all mesh reconstruction hypotheses

To overcome these limitations, we introduce MaskHand, a novel generative masked model designed for accurate 3D hand mesh recovery. By learning and sampling from the underlying joint distribution of hand articulations and image features, our model synthesizes most probable 3D hand meshes, mitigating ambiguities inherent in single-view reconstruction as shown in Figure 2. MaskHand consists of two main components: a VQ-MANO and a context-guided masked transformer, which are trained in two consecutive stages. In the first stage, VQ-MANO is trained with Vector Quantized Variational Autoencoders (VQ-VAE) [60] to encode continuous hand poses (e.g., joint rotations) of the MONO parametric hand model into a sequence of discrete pose tokens. The discretization in VQ-MANO is crucial for generative masked modeling, allowing the model to learn per-token probabilities conditioned on the input image. In the second stage, the token sequence is partially masked and the context-guided masked transformer is trained to reconstruct the masked tokens by learning token conditional distribution, based on multiple contextual clues, including corrupted token sequence, image features, and 2D pose structures. The learned distribution enables uncertainty quantification at each prediction step, facilitating confidence-guided sampling to iteratively refine mesh reconstructions. By selectively retaining high-confidence tokens during inference, MaskHand minimizes uncertainty in the 2D-to-3D mapping, enabling high-precision hand mesh recovery, even

in occluded and ambiguous scenarios.

- MaskHand is the first to leverage generative masked modeling for reconstructing robust 3D hand mesh. The main idea is to explicitly learn the 2D-to-3D probabilistic mapping and synthesize high-confidence, plausible 3D meshes by sampling from learned distribution.
- We design context-guided masked transformer that effectively fuses multiple contextual clues, including 2D pose, image features, and unmasked 3D pose tokens.
- We propose differential masked training to learn hand pose token distribution, conditioned on all contextual clues. This learned distribution facilitates confidence-guided sampling during inference, producing mesh reconstructions with low uncertainty and high precision.
- We show that MaskHand is a generic framework that unifies hand-mesh estimation and hand-mesh generation.
- We demonstrate through extensive experiments that MaskHand outperforms SOTA methods on standard datasets.

2. Related Work

2.1. Discriminative Methods

Human hand recovery has been developed in recent years, with early approaches [23, 29, 59, 62, 71, 73] leveraging optimization techniques to estimate hand poses based on 2D skeleton detections. Later, MANO [52] introduced a differentiable parametric mesh model that capture hand shape and articulation, allowing the model to provide a plausible mesh with end-to-end estimation of model parameters directly from a single-view image. Boukhayma et al. [6] presented the first fully learnable framework to directly predict the MANO hand model parameters [52] from RGB images. Similarly, several subsequent methods have leveraged heatmaps [70] and iterative refinement techniques [4] to ensure 2D alignment. Kulon et al. [35, 36] proposed a different regression approach that predicts 3D vertices instead of MANO pose parameters, achieving notable improvements over prior methods. Recent methods, such as METRO [49], MeshGraphormer [40], HaMeR [48] achieve the SOTA performance by modeling both vertex-vertex and vertex-joint interactions. These existing methods, based on discriminative regression, learn a deterministic mapping from the input image to the output mesh. This deterministic approach struggles to capture the uncertainties and ambiguities caused by hand self-occlusions, interactions with objects, and extreme poses or camera angles, resulting in unrealistic hand mesh reconstructions.

2.2. Generative Methods

Our MaskHand employs a generative method that learns a probabilistic mapping from the input image to the output mesh. It utilizes this learned distribution to synthesize high-confidence, plausible 3D hand meshes based on

2D visual contexts. HHMR [37] is the only other generative hand mesh recovery method in the literature. Unlike HHMR [37] that utilizes diffusion models, MaskHand is inspired by the success of masked image and language models for image and text generation tasks [8, 9, 18, 19, 72]. This fundamental difference allows MaskHand to explicitly and quantitatively estimate confidence levels or prediction probabilities for all mesh reconstruction hypotheses, enabling confidence-guided hypothesis selection for accurate reconstruction, as shown in Fig. 2. In contrast, HHMR’s denoising diffusion process synthesizes multiple mesh hypotheses without associating a confidence level with each hypothesis. Thus, it only reports the theoretically best mesh reconstruction by finding the hypothesis with minimal reconstruction errors under the assumption that ground-truth meshes are available.

3. Proposed Method: MaskHand

Problem Formulation. Given a single-hand image I , we aim to learn a mapping function $f(I) = \{\theta, \beta, \pi\}$ that output the MANO [52] model parameters from the input image. This mapping function encompasses three key components: the hand pose parameters $\theta \in \mathbb{R}^{48}$, shape parameters $\beta \in \mathbb{R}^{10}$, and camera parameters $\pi \in \mathbb{R}^3$, enabling comprehensive 3D hand reconstruction.

Overview of Proposed Method. As shown in Figure 3, MaskHand consists of two primary modules: VQ-MANO and the Context-Guided Masked Transformer. Initially, VQ-MANO discretizes 3D MANO pose parameters (θ) into discrete tokens, enabling generative masked modeling and uncertainty quantification. Next, the Context-Guided Masked Transformer leverages an encoder-decoder structure: the encoder extracts multi-scale image features, which, combined with 2D pose cues and unmasked tokens, are processed by the masked decoder. The decoder employs Graph-based Anatomical Pose Refinement to capture joint dependencies and a Context-Infused Masked Synthesizer to fuse visual features and token interactions, learning the probabilistic reconstruction via differential masked modeling. During inference, pose predictions are iteratively refined by selectively retaining high-confidence tokens and re-masking uncertain ones, progressively improving accuracy. Finally, the reconstruction is completed as the predicted pose (θ), shape (β) and camera parameters (π) are feed into the MANO hand model.

3.1. Hand Model and VQ-MANO

Our approach employs the MANO model [52], which maps pose $\theta \in \mathbb{R}^{48}$ and shape $\beta \in \mathbb{R}^{10}$ to a 3D hand mesh $M \in \mathbb{R}^{778 \times 3}$ and joint locations $X \in \mathbb{R}^{21 \times 3}$, representing both hand surface and pose.

VQ-MANO is a MANO hand tokenizer that employs a 1D convolutional autoencoder to learn a discrete latent space for 3D pose parameters $\theta \in \mathbb{R}^{48}$ by quantizing the

continuous pose embeddings into a learned codebook C with discrete code entries, as depicted in Figure 3(a). A key feature is its upsampling mechanism, expanding 16 MANO poses into 64 discrete pose tokens, which enhances spatial detail and improves mesh recovery accuracy (Supplementary Section C). To this end, we employ a Vector Quantized Variational Autoencoder (VQ-VAE) [60] for pretraining the tokenizer. Specifically, we input the MANO pose parameters θ into a convolutional encoder E , which maps them to a latent embedding z . Each embedding z_i is then quantized to its nearest codebook entry $c_k \in C$ based on Euclidean distance, defined as

$$\hat{z}_i = \arg \min_{c_k \in C} \|z_i - c_k\|_2.$$

The total loss function of VQ-MANO is formulated as:

$$\mathcal{L}_{\text{vq-mano}} = \lambda_{\text{re}} \mathcal{L}_{\text{recon}} + \lambda_{\text{E}} \|\text{sg}[z] - c\|_2 + \lambda_{\alpha} \|z - \text{sg}[c]\|_2,$$

where $\mathcal{L}_{\text{vq-mano}}$ consists of a MANO reconstruction loss $\mathcal{L}_{\text{recon}}$, a latent embedding loss, and a commitment loss, weighted by hyperparameters λ_{re} , λ_{E} , and λ_{α} , respectively. Here, $\text{sg}[\cdot]$ denotes the stop-gradient operator, which prevents gradients from flowing through its argument during backpropagation. To further enhance reconstruction quality, we incorporate an additional L1 loss:

$$\mathcal{L}_{\text{recon}} = \lambda_{\theta} \mathcal{L}_{\theta} + \lambda_{\text{V}} \mathcal{L}_{\text{V}} + \lambda_{\text{J}} \mathcal{L}_{\text{J}},$$

which aims to minimize the discrepancies between the predicted and ground-truth MANO parameters, including the pose parameters θ , mesh vertices V , and hand joints J . The tokenizer is optimized using a straight-through gradient estimator to facilitate gradient propagation through the non-differentiable quantization step. Additionally, the codebook entries C are updated via exponential moving averages and periodic codebook resets, as described in [21, 64].

3.2. Context-Guided Masked Transformer

The context-guided masked transformer comprises two main components: the multi-scale image encoder and the masked graph transformer decoder.

3.2.1. Multi-scale Image Encoder

Our encoder uses a vision transformer (ViT-H/16) [2] to extract image features [48], processing 16x16 pixel patches into feature tokens. Following ViTDet [38], we adopt a multi-scale feature approach by upsampling the initial feature map to produce feature maps at varying resolutions. This multi-scale representation is critical for handling complex articulations and self-occlusions in hand poses. High-resolution maps provide fine-grained joint details, while low-resolution maps capture global hand structure, balancing precision in joint positioning with overall anatomical coherence. Moreover, we utilize cross-attention with low-resolution feature maps in the x-Attention head to regress

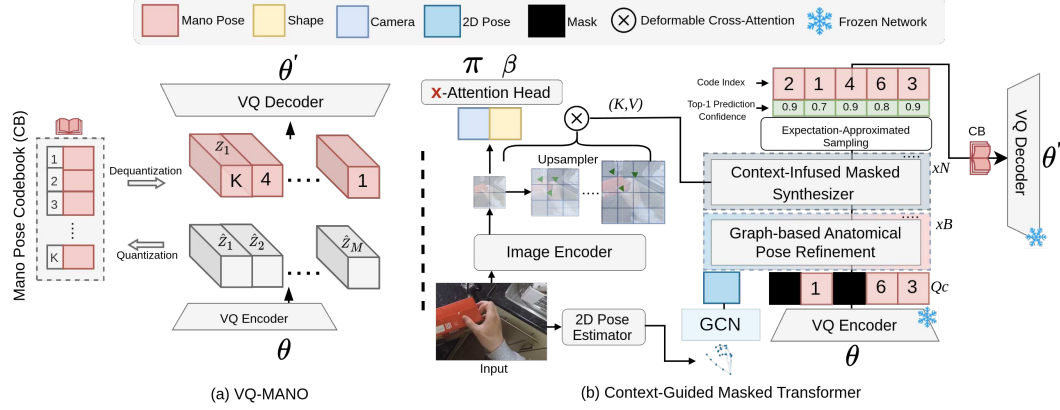


Figure 3. *MaskHand Training Phase*. MaskHand consists of two key components: (1) **VQ-MANO**, which encodes 3D hand poses into a sequence of discrete tokens within a latent space, and (2) a **Context-Guided Masked Transformer** that models the probabilistic distributions of these tokens, conditioned on the input image, 2D pose cues, and a partially masked token sequence.

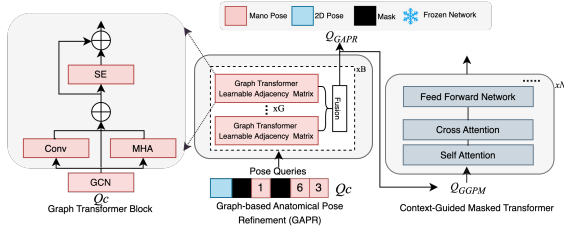


Figure 4. Architecture of Graph-based Anatomical Pose Refinement (GAPR) and Context-Infused Masked Synthesizer, illustrating fusion of anatomical pose dependencies and contextual cues for precise mesh reconstruction.

stable shape parameters (β) and camera orientation (π), making the process computationally efficient. This approach decouples shape estimation from pose modeling, preserving morphological stability and spatial alignment, and enhancing robustness and anatomical accuracy in 3D hand reconstruction.

3.2.2. Masked Graph Transformer Decoder

The Masked Transformer Decoder is composed of two key components (Figure 3 and 4): Graph-based Anatomical Pose Refinement (GAPR) and the Context-Infused Masked Synthesizer Module.

Graph-based Anatomical Pose Refinement (GAPR). Our decoder employs two blocks of a lightweight graph transformer that process pose tokens generated by VQ-MANO, enriched with 2D pose guidance, ensuring anatomically coherent hand reconstructions as shown in Figure 4. To construct a hand-kinematic aware 2D pose token, we first process OpenPose 2D keypoints [7] through a Graph Convolutional Network (GCN) with a fixed hand skeleton adjacency matrix, capturing explicit kinematic structures. Simultaneously, VQ-MANO pose tokens, which inherently

encode implicit and non-linear joint dependencies, are extracted. These two representations are then fused and refined through two graph transformer block, where learnable adjacency matrices adaptively model complex joint relationships, ensuring robust spatial and structural consistency. To further enhance stability and anatomical accuracy, we integrate a transformer encoder with a Squeeze-and-Excitation (SE) block [30], which emphasizes joint orientations and angles while ensuring spatial alignment through a 1×1 convolution layer. Within the transformer encoder, Multi-Head Attention (MHA) and pointwise convolution layers refine high-resolution dependencies between joints, producing a cohesive and anatomically aligned 3D pose representation, denoted as Q_{GAPR} . The output Q_{GAPR} is formulated as:

$$Q' = \text{MHA}(\text{Norm}(Q_C)) + \text{Conv}(\text{Norm}(Q_C)) + Q_C \quad (1)$$

$$Q_{GAPR} = \text{SE}(\text{Norm}(Q')) \quad (2)$$

where $Q_C \in \mathbb{R}^{K \times D}$ represents MANO pose tokens and 2D pose context queries, and Q_{GAPR} is the stabilized, anatomically consistent pose representation. These refined tokens Q_{GAPR} are then processed by the Context-Infused Masked Synthesizer to generate precise 3D hand meshes.

Context-Infused Masked Synthesizer. We leverage a multi-layer transformer (as shown in Figure 4) whose inputs are refined pose tokens Q_{GAPR} and cross-attends them with multi-scale feature maps generated by the image encoder. To enhance computational efficiency with high-resolution feature maps, a deformable cross-attention mechanism is employed [75]. This allows each pose token to focus on a selected set of sampling points around a learnable reference point, rather than the entire feature map. By concentrating attention on relevant areas, the model achieves a

balance between computational efficiency and spatial precision, preserving essential information for accurate 3D hand modeling. The deformable cross-attention is defined as:

$$\text{MCDA}(Q_{\text{GAPR}}, \hat{p}_y, \{x^l\}) = \sum_{l,k} A_{lyk} \cdot \mathbf{W} x^l(\hat{p}_y + \Delta p_{lyk})$$

where Q_{GAPR} are refined manopose token queries, $\hat{p}_y$ are learnable reference points, Δp_{lyk} are sampling offsets, $\{x^l\}$ are multi-scale features, A_{lyk} are attention weights, and $\mathbf{W}$ is a learnable weight matrix. With the inclusion of a [MASK] token in the masked transformer decoder, the module can predict masked pose tokens during training, while also facilitating token generation during inference. This approach uses the [MASK] token as a placeholder for pose predictions, enabling robust synthesis of occluded or unobserved hand parts for coherent 3D hand reconstruction.

3.3. Training: Differential Masked Modeling

Context-conditioned Masked Modeling. We employ masked modeling to train our model that learns the probabilistic distribution of 3D hand poses, conditioned on multiple contextual cues. Given a sequence of discrete pose tokens $Y = [y_i]_{i=1}^L$ from the pose tokenizer, where L is the sequence length, we randomly mask out a subset of m tokens with $m = \lceil \gamma(\tau) \cdot L \rceil$, where $\gamma(\tau)$ is a cosine-based masking ratio function. Here, τ is drawn from a uniform distribution $U(0, 1)$, and we adopt the masking function $\gamma(\tau) = \cos(\frac{\pi\tau}{2})$, inspired by generative text-to-image modeling strategies [8]. Masked tokens are replaced with learnable [MASK] tokens, forming a corrupted sequence $Y_{\overline{\mathbf{M}}}$ that the model must reconstruct. Each token y_i is predicted based on the probabilistic distribution $p(y_i | Y_{\overline{\mathbf{M}}}, X_{2D}, X_{img})$, conditioned on corrupted token sequence $Y_{\overline{\mathbf{M}}}$, 2D pose embedding X_{2D} , and image prompt X_{img} . This approach enables the model to explicitly account for the uncertainty inherent in mapping 2D observations to a coherent 3D hand mesh. The training objective is to minimize the negative log-likelihood of correctly predicting each pose token in the sequence, formulated as follows:

$$\mathcal{L}_{\text{mask}} = -\mathbb{E}_{Y \in \mathcal{D}} \left[\sum_{\forall i \in [1, L]} \log p(y_i | Y_{\overline{\mathbf{M}}}, X_{2D}, X_{img}) \right].$$

Expectation-Approximated Differential Sampling. The training objective $\mathcal{L}_{\text{mask}}$ captures stochastic uncertainty in hand mesh reconstruction, enabling precise estimation of pose parameters θ within a structured latent space. Recent studies [48] show that auxiliary 3D and 2D joint losses improve pose recovery by aligning predicted and ground-truth 2D/3D joints. Integrating these losses requires transforming latent pose tokens into θ , a process involving non-differentiable probabilistic sampling. To address this, instead of sampling the distribution $p(y_i | Y_{\overline{\mathbf{M}}}, X_{2D}, X_{img})$ to

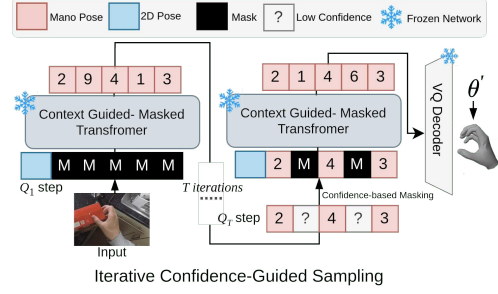


Figure 5. MaskHand Inference Phase: **Confidence-Guided Iterative Sampling** — a step-by-step refinement of pose selection by probabilistically sampling high-confidence tokens.

obtain the most probable token from codebook, we implement an expectation-based differential relaxation: instead of directly estimating discrete code indices, the model outputs logits L for each token. These logits undergo a softmax operation, producing the token distribution, which are multiplied by the pretrained codebook, resulting in the mean quantized feature representations $\bar{z} = [\bar{z}_1, \bar{z}_2, \dots, \bar{z}_M]$:

$$\bar{z} = \text{softmax}(L_{M \times K}) \times \text{CB}_{K \times D}$$

where L represents the logits matrix, CB denotes the codebook, M is the token count, K specifies the codebook size, and D defines the dimensionality of each codebook entry. This mean token embeddings $\bar{z}$ are feed into the decoder to reconstruct the pose parameter θ' . Combined with shape parameters β and camera parameters π , this process reconstructs the 3D hand mesh, enabling the computation of both 3D joint loss and 2D projection loss. Since obtaining $\bar{z}$ is differential, the model can be trained end-to-end using the overall loss function .

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{MANO}} + \mathcal{L}_{3D} + \mathcal{L}_{2D}$$

which combines the masked token prediction loss ($\mathcal{L}_{\text{mask}}$), 3D joint loss ($\mathcal{L}_{3D}$), 2D projection loss ($\mathcal{L}_{2D}$), and MANO parameter loss $\mathcal{L}_{\text{MANO}}$. Together, these terms collectively minimize discrepancies within MANO space, enabling precise 3D hand reconstruction.

3.4. Inference: Confidence-Guided Sampling

Our model leverages confidence-guided sampling to achieve precise and stable 3D hand pose predictions as shown in Figure 5. This process begins with a fully masked sequence Y_1 of length L , with each token initialized as [MASK]. Over T decoding iterations, each iteration t applies stochastic sampling to predict masked tokens based on their distributions $p(y_i | Y_{\overline{\mathbf{M}}}, X_{2D}, X_{img})$. Following each sampling step, tokens with the lowest prediction confidences are re-masked to be re-predicted in subsequent iterations. The number of tokens re-masked is determined by a

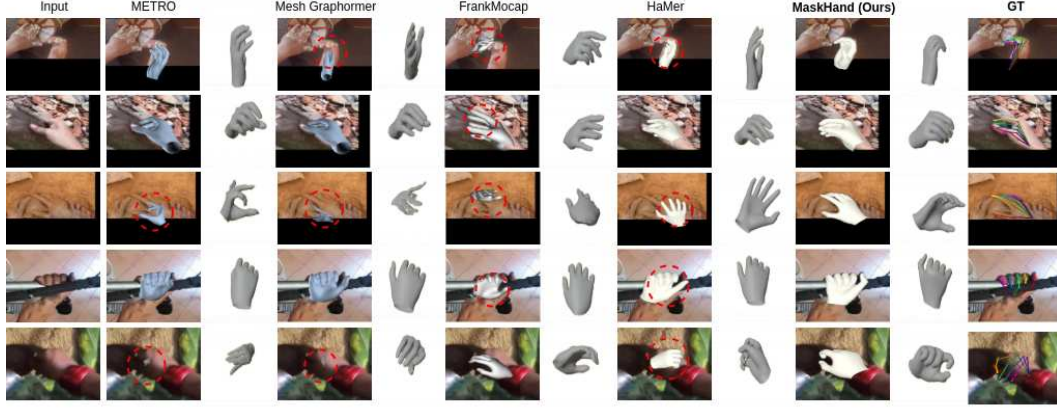


Figure 6. Comparison of SOTA methods: MaskHand synthesizes unobserved parts for accurate 3D reconstructions in occluded scenarios.

masking schedule $\lceil \gamma \left(\frac{t}{T} \right) \cdot L \rceil$, where γ is a decaying function of t . This schedule dynamically adjusts masking intensity based on confidence, using a higher masking ratio in earlier iterations when prediction confidence is lower. Consequently, the model iteratively refines ambiguous regions, progressively improving prediction confidence as context builds. The masking ratio decreases with each step, stabilized by the cosine decay function γ .

4. Experiments

Method	PA-MPJPE mm(↓)	PA-MPVPE mm(↓)	F@5 (↑)	F@15 (↑)	AUC _J (↑)	AUC _V (↑)
S ² HAND [13]	11.5	11.1	0.448	0.932	0.769	0.778
KPT-Transf. [28]	10.9	-	-	-	0.785	-
ArtiBoost [67]	10.8	10.4	0.507	0.946	0.785	0.792
Yu et al. [69]	10.8	10.4	-	-	-	-
HandGCAT [63]	9.3	9.1	0.552	0.956	0.814	0.818
AMVUR [32]	8.7	8.3	0.593	0.964	0.826	0.834
HMP [20]	10.1	-	-	-	-	-
SPMHand [42]	8.8	8.6	0.574	0.962	-	-
MaskHand	7.0	7.0	0.663	0.984	0.860	0.860

Table 1. Zero-Shot 3D Mesh Reconstruction Evaluation on HO3Dv3 [28] Benchmark: SOTA Comparison. ”-” indicates unreported metrics.

Training Datasets. To train the hand pose tokenizer, we employed a diverse set of datasets to capture a wide range of hand poses and interactions. This includes DexYCB [10], InterHand2.6M [45], MTC [65], and RHD [76]. For training MaskHand, we constructed a comprehensive dataset following a similar setup as [48] to ensure fair comparison. Specifically, for evaluations on HO3Dv3 [28] and HInt [48], MaskHand was trained on a diverse mix of datasets, including FreiHAND [77], HODv2 [27], MTC [65], RHD [76], InterHand2.6M [45], H2O3D [27], DexYCB [10], COCO-Wholebody [33], Halpe [22], and MPII NZSL [54]. However, for FreiHAND and DexYCB, MaskHand was trained specifically on their **respective training datasets**.

Methods	PA-MPJPE mm (↓)	PA-MPVPE mm (↓)	F@5 (↑)	F@15 (↑)
MeshGraphormer [†] [40]	5.9	6.0	0.764	0.986
FastMETRO [16]	6.5	7.1	0.687	0.983
FastViT [61]	6.6	6.7	0.722	0.981
AMVUR [32]	6.2	6.1	0.767	0.987
Deformer [68]	6.2	6.4	0.743	0.984
PointHMR [34]	6.1	6.6	0.720	0.984
Zhou et al. [74]	5.7	6.0	0.772	0.986
HaMeR [48]	6.0	5.7	0.785	0.990
HHMR [§] [37]	5.8	5.8	-	-
MaskHand[‡]	5.5	5.4	0.801	0.991

Table 2. 3D Mesh Reconstruction Accuracy on FreiHAND [77]: SOTA Comparison. [†] Denotes Test-Time Augmentation. PA-MPJPE/PA-MPVPE in mm. ”-” indicates unreported metrics.

[‡] **MaskHand** produces confident mesh reconstruction, guided by the learned 2D-to-3D distribution **without** assuming GT meshes.
[§] HHMR reports best hypothesis mesh reconstruction with minimum MPJPE and MPVPE, **with** the assumption of available GT meshes.

Method	PA-MPJPE (↓)	PA-MPVPE (↓)	MPJPE (↓)	MPVPE (↓)
METRO [70]	7.0	-	-	-
Spurr et al. [55]	6.8	-	-	-
Liu et al. [41]	6.6	-	-	-
HandOceNet [46]	5.8	5.5	14.0	13.1
MobRecon [11]	6.4	5.6	14.2	13.1
H2ONet [66]	5.7	5.5	14.0	13.0
Zhou et al. [74]	5.5	5.5	12.4	12.1
Ours	5.0	4.9	11.7	11.2

Table 3. 3D Mesh Reconstruction Accuracy on DexYCB [10]: SOTA Comparison. ”-” indicates unreported metrics.

Evaluation Metrics. We follow standard protocols [48] to assess 3D joint reconstruction using PA-MPJPE and AUC_J. For 3D mesh evaluation, we use PA-MPVPE, AUC_V, and F@5mm/F@15mm to measure vertex precision. Moreover, PCK evaluates 2D reconstruction by reprojecting 3D joints into 2D space. The computational efficiency is measured by Average Inference Time per Image (AITI) in seconds [7], where lower values indicate better performance.

Method	Venue	NewDays			VISOR			Ego4D		
		@0.05 (↑)	@0.1 (↑)	@0.15 (↑)	@0.05 (↑)	@0.1 (↑)	@0.15 (↑)	@0.05 (↑)	@0.1 (↑)	@0.15 (↑)
Occluded Joints Evaluation										
FrankMocap [53]	ICCVW 2021	9.2	28.0	46.9	11.0	33.0	55.0	8.4	26.9	45.1
METRO [39]	CVPR 2021	7.0	23.6	42.4	10.2	32.4	53.9	8.0	26.2	44.7
MeshGraphormer [40]	ICCV 2021	7.9	25.7	44.3	10.9	33.3	54.1	9.3	32.6	51.7
HandOccNet (param) [47]	CVPR 2022	7.2	23.5	42.4	7.4	26.1	46.7	7.2	26.1	45.7
HandOccNet (no param) [47]	CVPR 2022	9.8	31.2	50.8	9.9	33.7	55.4	9.6	31.1	52.7
HaMeR [48]	CVPR 2024	27.2	60.8	78.9	25.9	60.8	80.7	23.0	56.9	76.3
MaskHand	Ours	29.4	64.1	80.3	31.4	64.2	83.6	29.4	65.3	80.1
Occluded Joints Evaluation with 80% Masked Hand										
HaMeR [48]	CVPR 2024	8.7	28.4	48.9	7.1	24.4	41.7	7.1	23.4	40.2
MaskHand	Ours	10.5	31.6	52.0	8.2	25.0	41.9	8.2	26.2	44.1
Occluded Joints Evaluation with 90% Masked Hand										
HaMeR [48]	CVPR 2024	8.7	27.4	48.1	7.1	24.3	42.2	7.1	23.4	40.2
MaskHand	Ours	10.4	31.0	51.1	9.1	27.3	44.3	8.0	25.1	43.2

Table 4. Zero-shot evaluation on the HInt Benchmark [48] for occluded joint reconstruction using the PCK metric, comparing MaskHand to SOTA methods. MaskHand demonstrates superior robustness and resilience to missing hand structures under severe occlusions, maintaining strong performance even with 80% and 90% masked hand regions. **None of the models were trained on the HInt dataset.**

Reconstruction Evaluation. We evaluate MaskHand’s joint and mesh reconstruction on HO3Dv3 [28], DexYCB [10], FreiHAND [77], and HInt [48] benchmarks.

Zero-Shot Generalization Evaluation: The HO3Dv3 dataset, with 20K annotated images, serves as a benchmark for evaluating generalization in complex hand-object interactions. MaskHand is tested on HO3Dv3 **without prior training** to assess its capability in handling unseen data, occlusions, dataset biases, and diverse hand articulations in real-world scenarios. Additionally, we evaluate MaskHand’s zero-shot performance on the HInt benchmark [48], which presents challenges such as varied lighting, extreme viewpoints, and complex hand-object interactions. Using the PCK metric, MaskHand is tested on HInt-NewDays [14], HInt-EpicKitchensVISOR [17], and HInt-Ego4D [26], demonstrating its ability to operate in diverse and occlusion-heavy environments.

Supervised Evaluation: We evaluated MaskHand on FreiHAND [77], which consists of 4K images captured in controlled environments, and DexYCB [10], a large-scale dataset containing 78K images of hands interacting with objects. For a fair comparison, we trained MaskHand exclusively on FreiHAND, following the evaluation protocols of MeshGraphormer [40] and HHMR [37]. Similarly, aligning with recent work [74], we trained MaskHand on DexYCB and conducted evaluations under the same conditions, ensuring a direct and consistent comparison.

4.1. Comparison to State-of-the-art Approaches

We evaluate MaskHand on HO3Dv3, FreiHAND, DexYCB, and HInt (Tables 1–4), where it outperforms SOTA methods, demonstrating strong zero-shot generalization and in-the-wild robustness. Unlike prior approaches, MaskHand leverages confidence-aware masked modeling to effectively handle uncertainty, occlusions, and complex hand poses. In zero-shot evaluation on HO3Dv3 (Table 1),

MaskHand reduces PA-MPJPE by 19.5% and PA-MPVPE by 15.7%, demonstrating superior dataset-agnostic reconstruction. On FreiHAND (Table 2), it achieves 3.5% lower PA-MPJPE and 5.3% lower PA-MPVPE, with the highest F@5mm (0.801) and F@15mm (0.991), ensuring precise, high-fidelity hand recovery. In DexYCB (Table 3), MaskHand reduces MPJPE by 9.1% and MPVPE by 10.9%, confirming its robustness in real-world hand-object interactions. In zero-shot evaluation on HInt (Table 4), which includes challenging egocentric, heavily occluded hand poses, MaskHand achieves state-of-the-art occluded joint reconstruction, improving PCK@0.05 by 8.1% on HInt-NewDays, 21.2% on HInt-VISOR, and 27.8% on HInt-Ego4D. Qualitatively, Figure 6 shows MaskHand’s superior synthesis of occluded regions compared to prior methods, while Figure 7 demonstrates its robustness under extreme masking. These results highlight MaskHand’s strong generalization and practical effectiveness in challenging real-world conditions.

4.2. Ablation Study

The key to MaskHand’s effectiveness lies in its generative masked modeling and uncertainty quantification, enabling robust reconstruction in occluded regions. Its iterative decoding strategy refines predictions, progressively enhancing accuracy in challenging scenarios. This ablation study explores iterative refinement and MaskHand’s versatility in text-conditioned 3D hand synthesis.

Further analyses, experimental details, and qualitative visualizations in the Supplementary Material explore key aspects, including the influence of *architectural choices* (e.g., VQ-MANO tokenizer, proposed modules, multi-scale feature resolutions, deformable attention layers), *training strategies* (masking ratios, regularization losses via differential sampling), and advanced *MaskHand capabilities* such as unconditional mesh generation and robust recon-



Figure 7. SOTA Comparison: Qualitative zero-shot evaluation on HInt Benchmark [48] for heavily masked hand images. MaskHand reconstructs occluded hand poses, demonstrating robustness to severe occlusions and generalizability to unseen masked regions.

struction under severe real-world occlusions.

Impact of Iterative Confidence-Guided Sampling. In confidence-guided sampling, we iteratively refine predictions by retaining high-confidence tokens and re-masking low-confidence ones, progressively enhancing stability and accuracy. As shown in Table 5, increasing iterations from 1 to 5 reduces PA-MPVPE (7.2 $\rightarrow$ 7.0 on HO3Dv3, 5.6 $\rightarrow$ 5.4 on FreiHAND). These improvements result from confidence-guided masking of uncertain regions during decoding, enhancing reconstruction accuracy.

# of iter.	HO3Dv3		FreiHAND		AITI (s) (↓)
	PA-MPJPE	PA-MPVPE	PA-MPJPE	PA-MPVPE	
1	7.2	7.2	5.6	5.6	0.04
3	7.1	7.1	5.6	5.5	0.08
5	7.0	7.0	5.5	5.4	0.12

Table 5. Iterations in Iterative Confidence-Guided Sampling. AITI is obtained on a single mid-grade GPU (NVIDIA RTX A5000).

4.3. Application: Text-to-Mesh Generation

MaskHand’s modular design extends beyond hand mesh reconstruction to text-conditioned 3D hand synthesis. By replacing the image encoder with a CLIP-based text encoder [51], MaskHand can generate 3D hand meshes from text inputs, as shown in Figure 8. Using the ASL dataset [1], we trained MaskHand with pseudo-ground-truth annotations generated by MaskHand. During testing, a top-5% probabilistic sampling strategy ensured diverse yet plausible mesh outputs, evaluated using Hausdorff Distance, Chamfer Distance, and PA-MPVPE. Results (Table 6) demonstrate high fidelity and consistency in text-driven reconstruction, with qualitative examples shown in Figure 9. More details are provided in Supplementary Section B.

Metric	Mean	Standard Deviation
Hausdorff Distance	0.0221	0.0073
Chamfer Distance	9.73×10^{-5}	5.47×10^{-5}
PA-MPVPE (mm)	12.2	3.1

Table 6. Quantitative evaluation of text-to-mesh generation using MaskHand’s Masked Synthesizer, reporting mean and standard deviation across all generated meshes per text prompt.

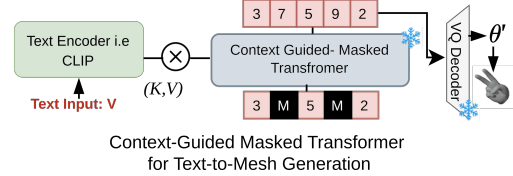


Figure 8. Repurposed MaskHand for Text-to-Mesh Generation

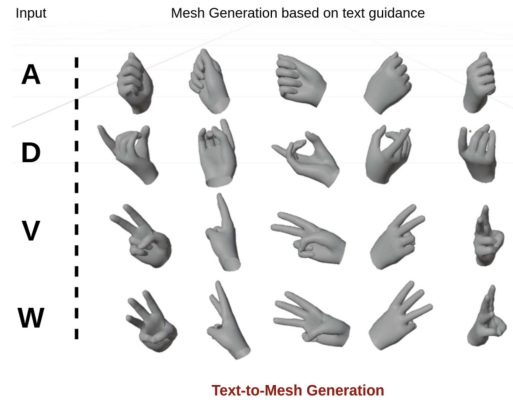


Figure 9. Text-to-mesh generation results on the ASL dataset, demonstrating MaskHand’s Masked Synthesizer’s ability to generate structurally accurate and diverse meshes from text inputs.

5. Conclusion

We introduced MaskHand, a generative masked model for accurate and robust 3D hand mesh reconstruction, particularly in occluded and ambiguous regions. By explicitly learning and sampling from the probabilistic 2D-to-3D mapping, MaskHand effectively handles self-occlusions, depth ambiguities, and missing hand parts. At its core, VQ-MANO encodes 3D hand articulations as discrete pose tokens, enabling structured latent representations. The Context-Guided Masked Transformer models joint token dependencies by incorporating image context, 2D pose cues, and masked token distributions. Through confidence-guided iterative sampling, MaskHand progressively refines reconstructions, selectively updating uncertain regions to produce anatomically coherent and high-precision hand meshes. MaskHand outperforms state-of-the-art methods in occlusion-heavy scenarios, setting a new benchmark for 3D hand modeling in HCI, AR, and VR.

References

- [1] Kaggle asl alphabet. Available online: <https://www.kaggle.com/grassknoted/asl-alphabet> (accessed on 19 July 2021). 8
- [2] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3
- [3] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Augmented skeleton space transfer for depth-based hand pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1
- [4] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 2
- [5] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *European Conference on Computer Vision (ECCV)*, 2012. 1
- [6] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10843–10852, 2019. 2
- [7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 4, 6
- [8] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 3, 5
- [9] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 3
- [10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 6, 7
- [11] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyan Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20544–20554, 2022. 6
- [12] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1
- [13] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10451–10460, 2021. 6
- [14] Tianyi Cheng, Dandan Shan, Ayda Hassen, Richard Higgins, and David Fouhey. Towards a richer 2d understanding of hands at scale. *Advances in Neural Information Processing Systems*, 36:30453–30465, 2023. 7
- [15] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision (ECCV)*, 2022. 2
- [16] Youwang Kim Oh Tae-Hyun Cho, Junhyeong. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision*, pages 342–359. Springer, 2022. 6
- [17] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 7
- [18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arxiv. arXiv preprint arXiv:1810.04805*, 2019. 3
- [19] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 3
- [20] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J Black. Hmp: Hand motion priors for pose and shape estimation from video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6353–6363, 2024. 6
- [21] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3
- [22] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alpha-pose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173, 2022. 6
- [23] Chengying Gao, Yujia Yang, and Wensheng Li. 3d interacting hand pose and shape estimation from a single rgb image. *Neurocomputing*, 2022. 2
- [24] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1
- [25] Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla. Large-scale multiview 3d hand pose dataset. *arXiv preprint arXiv:1707.03742*, 2017. 1
- [26] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson

- Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 7
- [27] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent 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*, pages 3196–3206, 2020. 6
- [28] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent 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*, pages 11090–11100, 2022. 6, 7
- [29] Ludovic Hoyet, Kenneth Ryall, Rachel McDonnell, and Carol O’Sullivan. Sleight of hand: perception of finger motion from reduced marker sets. In *SIGGRAPH*, 2012. 2
- [30] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4
- [31] Rongtian Huo, Qing Gao, Jing Qi, and Zhaojie Ju. 3d human pose estimation in video for human-computer/robot interaction. In *Intelligent Robotics and Applications*, pages 176–187, Singapore, 2023. Springer Nature Singapore. 1
- [32] Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan 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*, pages 758–767, 2023. 6
- [33] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 196–214. Springer, 2020. 6
- [34] Jeonghwan Kim, Mi-Gyeong Gwon, Hyunwoo Park, Hyukmin Kwon, Gi-Mun Um, and Wonjun Kim. Sampling is matter: Point-guided 3d human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 12880–12889, 2023. 6
- [35] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael Bronstein, and Stefanos Zafeiriou. Single image 3d hand reconstruction with mesh convolutions. *arXiv preprint arXiv:1905.01326*, 2019. 2
- [36] Dominik Kulon, Riza Alp Güler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4990–5000, 2020. 2
- [37] Mengcheng Li, Hongwen Zhang, Yuxiang Zhang, Ruizhi Shao, Tao Yu, and Yebin Liu. Hhmr: Holistic hand mesh recovery by enhancing the multimodal controllability of graph diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 645–654, 2024. 3, 6, 7
- [38] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022. 3
- [39] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 7
- [40] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021. 2, 6, 7
- [41] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong 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 (CVPR)*, pages 14687–14697, 2021. 6
- [42] Haofan Lu, Shuiping Gou, and Ruimin Li. Spmhand: Segmentation-guided progressive multi-path 3d hand pose and shape estimation. *IEEE Transactions on Multimedia*, 2024. 6
- [43] Anders Markussen, Mikkel Rønne Jakobsen, and Kasper Hornbæk. Vulture: a mid-air word-gesture keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 1073–1082, New York, NY, USA, 2014. Association for Computing Machinery. 1
- [44] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1
- [45] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020. 6
- [46] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handocnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6
- [47] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handocnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1496–1505, 2022. 7
- [48] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 2, 3, 5, 6, 7, 8
- [49] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12242–12254, 2025. 2

- [50] Chen Qian, Xiao Sun, Yichen Wei, Xiaou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1
- [51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 8
- [52] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2, 3
- [53] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1749–1759, 2021. 7
- [54] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 1, 6
- [55] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European conference on computer vision*, pages 211–228. Springer, 2020. 6
- [56] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 1
- [57] Srinath Sridhar, Anna Maria Feit, Christian Theobalt, and Antti Oulasvirta. Investigating the dexterity of multi-finger input for mid-air text entry. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, page 3643–3652, New York, NY, USA, 2015. Association for Computing Machinery. 1
- [58] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1
- [59] Jonathan Taylor, Vladimir Tankovich, Danhang Tang, Cem Keskin, David Kim, Philip Davidson, Adarsh Kowdle, and Shahram Izadi. Articulated distance fields for ultra-fast tracking of hands interacting. *TOG*, 2017. 2
- [60] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3
- [61] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5785–5795, 2023. 6
- [62] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *TOG*, 2020. 2
- [63] Shuaibing Wang, Shunli Wang, Dingkan Yang, Mingcheng Li, Ziyun Qian, Liuzhen Su, and Lihua Zhang. Handgcat: Occlusion-robust 3d hand mesh reconstruction from monocular images. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2495–2500. IEEE, 2023. 6
- [64] Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, and John Hughes. Hierarchical quantized autoencoders. *Advances in Neural Information Processing Systems*, 33:4524–4535, 2020. 3
- [65] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019. 6
- [66] Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing 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*, pages 17048–17058, 2023. 6
- [67] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu 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*, pages 2750–2760, 2022. 6
- [68] Yusuke Yoshiyasu. Deformable mesh transformer for 3d human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2023. 6
- [69] Ziwei Yu, Linlin Yang, You Xie, Ping Chen, and Angela Yao. Uv-based 3d hand-object reconstruction with grasp optimization. *arXiv preprint arXiv:2211.13429*, 2022. 6
- [70] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019. 2, 6
- [71] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Light-weight multi-person total capture using sparse multi-view cameras. In *ICCV*, 2021. 2
- [72] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. Ufcbert: Unifying multi-modal controls for conditional image synthesis. *Advances in Neural Information Processing Systems*, 34:27196–27208, 2021. 3
- [73] Wenping Zhao, Jinxiang Chai, and Ying-Qing Xu. Combining marker-based mocap and rgb-d camera for acquiring high-fidelity hand motion data. In *SIGGRAPH*, 2012. 2
- [74] Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2024. 6, 7
- [75] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ArXiv*, abs/2010.04159, 2020. 4

- [76] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. [6](#)
- [77] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. [6](#), [7](#)