

OASIS: Occlusion-aware Single-image Hand Avatar Reconstruction via 3D Gaussian Splatting

Zhisheng Han
University of Leicester
Leicester, UK
zh174@leicester.ac.uk

Shiyao Wu
University of Leicester
Leicester, UK
sw658@leicester.ac.uk

Jiayan Qiu
University of Leicester
Leicester, UK
jiayan.qiu.1991@outlook.com

Yakun Ju
University of Leicester
Leicester, UK
kelvin.yakun.ju@gmail.com

Lu Liu
University of Exeter
Exeter, UK
l.liu3@exeter.ac.uk

Le Zhang
University of Birmingham
Birmingham, UK
l.zhang.16@bham.ac.uk

Pengfei Feng
China University of Geoscience
Wuhan, China
20161004188@cug.edu.cn

Huiyu Zhou
University of Leicester
Leicester, UK
hz143@leicester.ac.uk

Zheheng Jiang[✉]
University of Leicester
Leicester, UK
zhehengjiangcareer@gmail.com

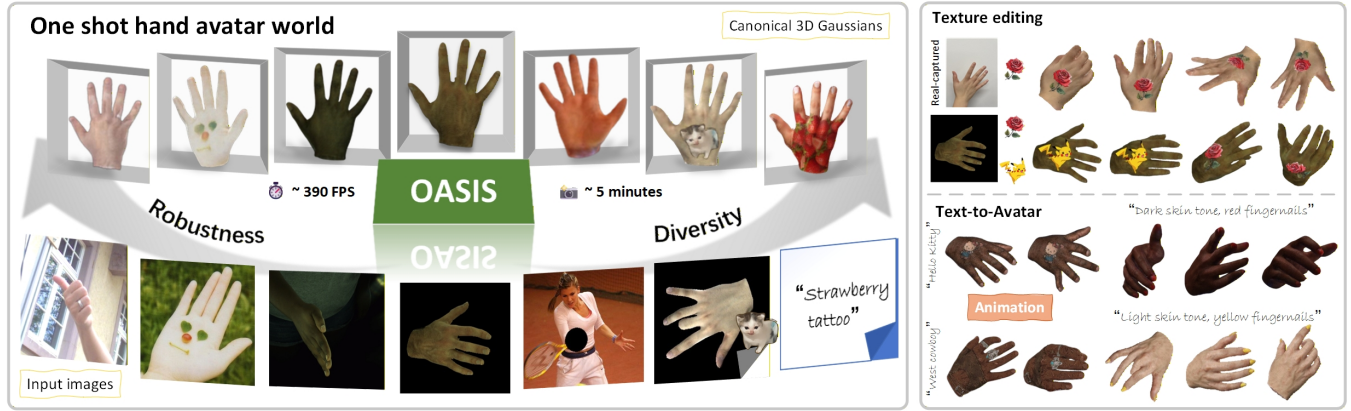


Figure 1: We introduce *OASIS*, a 3DGS-based one-shot hand avatar reconstruction model with efficient adaptation (~5 minutes) and real-time rendering (~390 FPS). We showcase a gallery of one-shot hand avatars that highlights both robustness and diversity: it reconstructs high-fidelity hands under challenging poses, complex textures, and extreme side-view observations, while also supporting human-image inputs and downstream applications such as texture editing and text-to-avatar generation.

Abstract

Single-image 3D hand avatar reconstruction is fundamentally ill-posed and particularly challenging due to limited visual evidence under severe self-occlusion and the complex pose-dependent deformation of highly articulated hands. Existing methods predominantly rely on implicit NeRF-style representations, whose volumetric fitting is computationally expensive and often struggles to preserve fine-grained hand details. In this work, we present *OASIS*, a tailored 3D Gaussian Splatting framework for single-image hand avatar reconstruction. To faithfully encode sparse image-specific appearance

cues in single-view reconstruction, we construct geometry-aligned visual evidence tokens by explicitly aligning input image observations with 3D hand geometry and context-adaptively tokenizing the resulting visual evidence. Since severe self-occlusion makes the reliability of image evidence inherently visibility-dependent, we introduce a visibility-conditioned point-image attention to reliably transfer visual evidence to geometric tokens, yielding occlusion-aware Gaussian features for faithful and robust reconstruction. To further capture non-rigid deformation of articulated hands, we introduce a Feature-on-Mesh representation to enable Gaussian deformation to be guided by local surface stretching. Under this framework, we adopt a one-shot adaptation scheme that learns a shared hand prior from multi-identity training data and then fits it to a target image for target-specific reconstruction. Extensive experiments show that *OASIS* outperforms existing baselines in both visual fidelity and efficiency across challenging poses and



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
<https://doi.org/10.1145/3767308.3835367>

in-the-wild scenarios, and further demonstrates strong versatility in downstream applications such as text-to-avatar generation and texture editing. Code will be released on the [project page](#).

CCS Concepts

• **Computing methodologies** → **Appearance and texture representations**; *Reconstruction*; 3D imaging.

Keywords

Hand avatar; Single image reconstruction; 3D Gaussian Splatting.

ACM Reference Format:

Zhisheng Han, Shiyao Wu, Jiayan Qiu, Yakun Ju, Lu Liu, Le Zhang, Pengfei Feng, Huiyu Zhou, and Zheheng Jiang. 2026. OASIS: Occlusion-aware Single-image Hand Avatar Reconstruction via 3D Gaussian Splatting. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil*. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3767308.3835367>

1 Introduction

Single-image 3D avatar reconstruction is an emerging modeling technique that aims to create a life-like 3D avatar from just one image, offering a highly cost-effective and promising approach to 3D modeling [27, 28]. Recent advances have made this direction increasingly attractive for realistic digital human modeling [1, 27, 28, 35, 38]. Among different avatar types, hand avatars are of particular importance because hands are a primary medium for interacting with the physical world through grasping and manipulation, while also providing rich expressive cues for gesture-based communication in virtual environments [5, 13, 17, 32].

Despite its practical importance, hand avatar reconstruction remains relatively underexplored compared with full-body avatar modeling. Prior hand avatar methods [3, 4, 6] have largely relied on Neural Radiance Fields (NeRF) [16] and related implicit neural fields. However, implicit volumetric rendering remains a practical bottleneck for avatar applications, as dense ray sampling and repeated field evaluations are computationally expensive, limiting interactivity and practical deployment [26, 31]. To address this broader limitation of implicit avatar rendering, recent human research [9, 27, 28, 35, 38, 39] has increasingly adopted explicit 3D Gaussian Splatting (3DGS) [12] representation, which enables efficient rendering while maintaining photorealistic quality. Nevertheless, exploiting these advantages in hand avatar reconstruction remains non-trivial.

Recent 3DGS-based avatar methods achieve high reconstruction fidelity and rendering efficiency for animatable human avatars. However, extending them to single-view hand reconstruction is non-trivial, as hands exhibit significantly greater articulation and self-occlusion than body or head avatars, making single-image visual evidence sparse and ambiguous. To address this, we construct geometry-aligned visual evidence tokens that preserve image-specific appearance by aligning it with 3D hand geometry and encoding point-aligned observations into compact tokens. To address the challenge that self-occlusion deprives occluded points of reliable observations and makes them more susceptible to interference from other surfaces, we propose a visibility-conditioned

point-image attention module that adaptively modulates the integration of image appearance cues into geometric tokens based on visibility. To further improve non-rigid hand modeling by capturing local surface stretching, we introduce a Feature-on-Mesh representation that organizes fused features over mesh faces. This design enables the decoded Gaussians to deform coherently with mesh-face stretching under articulation, yielding smoother and more consistent appearance.

Through training on accessible multi-identity data, the proposed single-image reconstruction framework learns prior reconstruction knowledge. As discussed in [41], one-shot hand avatar reconstruction requires target-specific personalization to adapt a prior model to the input image—capturing observed appearance while relying on the prior for unseen regions. Our goal is to inject identity-specific appearance and fine texture details into the Gaussian representation while preserving priors in unobserved areas. To this end, we first apply color calibration to compensate for global appearance discrepancies between the input image and the prior. We then fine-tune the Gaussian decoder with LoRA [8] to capture target-specific textures. This personalization enables robust identity adaptation and high-fidelity reconstruction for one-shot hand avatars. Our main contributions are:

- We introduce OASIS, a 3DGS-based one-shot hand avatar reconstruction model for reconstructing animatable and high-fidelity hand avatars from a single image.
- In OASIS, we first construct geometry-aligned visual evidence tokens and perform visibility-conditioned evidence transfer to produce occlusion-aware Gaussian features, enabling faithful and robust single-view hand reconstruction. We further introduce a Feature-on-Mesh representation to better capture local surface deformation under non-rigid hand articulation.
- Extensive experiments on InterHand2.6M and in-the-wild images demonstrate that our method consistently outperforms existing approaches in fidelity, robustness, and rendering efficiency. We further show the versatility of our pipeline on text-to-avatar generation and texture editing.

2 Related Work

Single-image Animatable Gaussian Avatar Reconstruction. Reconstructing an animatable 3D avatar from a single image is fundamentally ill-posed due to incomplete observations and ambiguous geometry. Recent 3DGS-based methods alleviate this difficulty mainly through prior-guided synthesized observations and prior-driven reconstruction models. SinGS [35] leverages Kinematic Human Diffusion to synthesize pose-space auxiliary observations with high 3D consistency for single-image reconstruction, and further combines compact 3D distillation to consolidate incomplete monocular evidence into Gaussian avatars. AniGS [28] synthesizes detailed multi-view canonical-pose observations with a pretrained generative model to reduce reconstruction ambiguity, and addresses their cross-view inconsistency through a 4D Gaussian formulation. LHM [27] adopts a large multimodal reconstruction model that fuses image and geometric features for direct Gaussian avatar prediction. LAM [7] further exploits a canonical-space parametric prior to enable direct animatable Gaussian reconstruction in the head

domain. Among these works, our approach is most closely related to prior-driven reconstruction models, which are better suitable for hands as they directly predict canonical-space Gaussians. By contrast, methods relying on auxiliary observations are less practical, as the strong hand articulation make such observations harder to synthesize consistently and accurately. Distinct from the human body and head, the hand is a highly articulated structure [3], making single-image reconstruction more challenging due to severe self-occlusion, ambiguous monocular evidence, and fine-grained non-rigid deformation. These hand-specific challenges are not fully captured by existing 3DGS avatar methods. To address this gap, we propose a hand-tailored one-shot 3DGS reconstruction model.

Animatable Hand Avatar. Existing research on animatable hand avatars mainly follows two lines. Early methods are predominantly based on explicit parametric hand models with texture representations. HTML [25] introduces a parametric texture space on top of MANO model [29]. Building on this line, NIMBLE [15] and Handy [22] further improve surface and texture realism through richer hand modeling and learned texture priors. HARP [11] advances this direction by adapting both geometry and appearance to better capture identity-specific characteristics. Nevertheless, their visual fidelity is still limited by the coarse geometry and sparsity of parametric hand meshes. More broadly, advances in neural avatars, such as [34] and [10], have demonstrated the potential of neural representations for modeling articulated appearance. However, directly extending such full-body formulations to hands remains challenging due to the more severe self-contact, self-occlusion, and fine-grained appearance variations, motivating hand-specific methods for realistic reconstruction. LISA [4] is the first to model hand shape and appearance with NeRF, while HandAvatar [3] further disentangles hand geometry, albedo, and illumination for high-fidelity animatable hand avatars. Subsequent methods such as HandNeRF [6] and LiveHand [19] improve deformation-aware neural rendering and real-time performance under hand motion. However, these methods typically rely on multi-view images or monocular videos. OHTA [41] begins to explore one-shot hand avatar reconstruction from a single image by learning transferable hand priors and performing test-time inversion and fitting. However, it still builds on implicit volumetric representations, which are computationally expensive to optimize and render, and shows limited ability to faithfully reconstruct high-frequency details due to the indirect coupling between image observations and 3D structure, thereby limiting practical usability. In contrast, our work develops a 3DGS-based one-shot hand avatar framework, enabling efficient rendering together with faithful reconstruction from a single image.

3 Method

3.1 Overview

Given a single RGB hand image $I \in \mathbb{R}^{H \times W \times 3}$, our goal is to reconstruct an animatable 3D hand avatar represented by 3D Gaussians. To this end, we propose OASIS, an animatable Gaussian hand reconstruction model, as shown in fig. 2. Our method addresses three key challenges in single-image hand reconstruction: preserving sparse image-specific appearance evidence, robustly transferring such evidence to 3D representations under severe self-occlusion,

and modeling local surface Gaussian deformation under hand articulation. Specifically, we design a geometry-aligned visual token construction module (Sec. 3.2) that extracts geometry-aligned image evidence and encodes it into a compact set of visual evidence tokens (VETs). Built upon the geometry-aligned VETs, we further develop a visibility-conditioned point-image attention (Sec. 3.3) module to reliably transfer visual evidence to geometric tokens under severe self-occlusion, producing occlusion-aware Gaussian features for robust hand reconstruction. Moreover, we introduce a Feature-on-Mesh (FoM) representation (Sec. 3.4) over mesh faces, enabling Gaussian deformation to be guided by local stretching. Finally, we decode the fused features into a canonical set of per-Gaussian attributes, including position, scaling, rotation, opacity, and color, and animate them to target poses via linear blend skinning (LBS).

3.2 Geometry-aligned Visual Token Construction

In single-image hand avatar reconstruction, a central challenge lies in effectively lifting 2D appearance observations into 3D space for animatable representation. Prior work [41] relies on an implicit volumetric paradigm, which is computationally expensive and only indirectly propagates 2D image evidence to 3D structure, making fine-grained and robust reconstruction difficult under severe self-occlusion and articulation. Recent human works [27, 36, 38] instead adopt an explicit 3DGS reconstruction pipeline, typically using backbone image features and transferring them to geometric tokens. However, such generic image-feature transfer is insufficient for hands, where limited single-view evidence is further degraded by severe self-occlusion, making reliable association between local image observations and the 3D hand geometry substantially more difficult. Therefore, we introduce a geometry-guided projection to explicitly establish geometry-consistent correspondence, which aligns backbone image features with posed 3D points, converting image-domain representations into point-aligned visual observations. Building on these projected image features, we further propose a learnable tokenization scheme that content-adaptively encodes them into informative yet compact visual evidence tokens, reducing redundancy caused by nearby projections.

3.2.1 Geometry-aligned Feature Projection and aggregation. Given a source image, we extract dense image features using a DINOv2 backbone [21], obtaining a feature map defined in the image domain $F \in \mathbb{R}^{h \times w \times C}$, where $h \times w$ denotes the spatial resolution of the feature map and C is the feature dimension. In parallel, we make the initialization of 3D points $\{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^{N_{\text{point}}}$ relying on the hand prior model MANO-HD [3], a super-resolution version of hand model MANO [29], to maintain the hand topology. Given the MANO shape and pose parameters estimated from the input image using an off-the-shelf estimator [23], these points could be deformed to posed space via LBS, resulting in posed 3D points. Instead of directly using backbone image features as in [7, 27, 36], we propose to align image features with posed hand geometry to make explicit 2D-3D correspondence for faithful reconstruction under limited visual cues. Specifically, we first project each posed point onto the image plane under the given camera extrinsic matrices: $\mathbf{p}_i^{\text{cam}} = \mathbf{R} \mathbf{p}_i + \mathbf{T}$, where $\mathbf{R} \in \mathbb{R}^{3 \times 3}$ and $\mathbf{T} \in \mathbb{R}^3$ denote the camera rotation and translation,

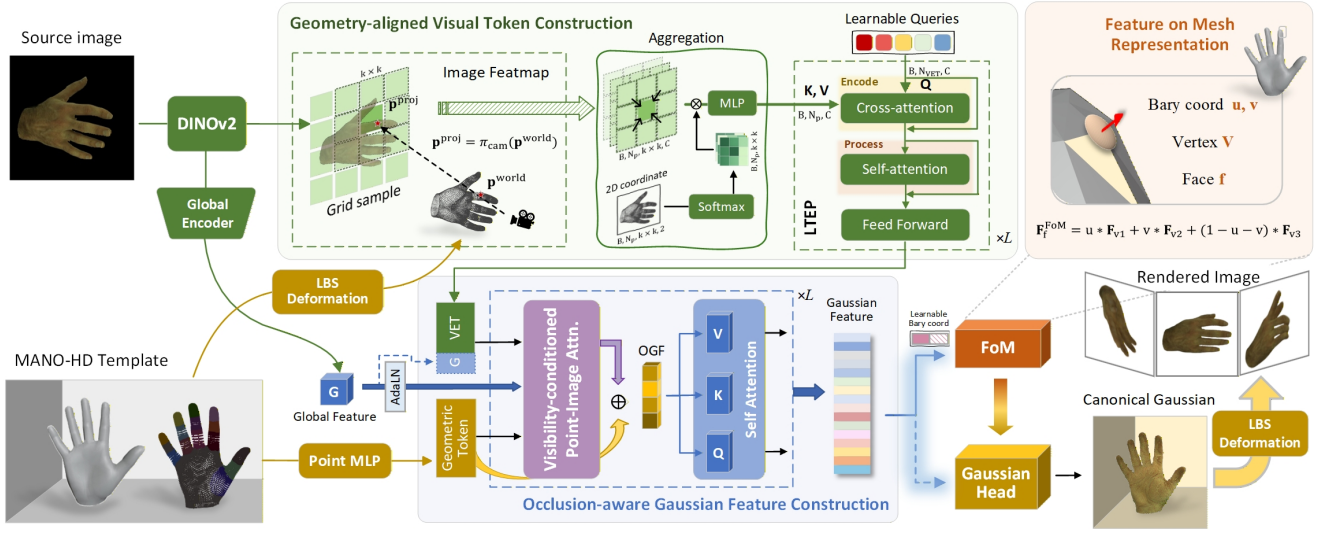


Figure 2: Overview of the proposed OASIS. Given a single input image, we construct geometry-aligned VETs to encode image-specific appearance cues, and use visibility-conditioned point-image cross-attention to transfer visual evidence to geometry-anchored point tokens, yielding occlusion-aware Gaussian features (OGF) under self-occlusion. We then refine these features with self-attention and lift them onto mesh faces via the FoM representation, providing a face-level carrier for Gaussian deformation under non-rigid hand animation. The final Gaussian features are decoded into Gaussian attributes in canonical space, followed by animating to target poses via LBS.

respectively. The camera-space points are then projected onto the image plane using the camera intrinsic matrix $\mathbf{K}$: $\tilde{\mathbf{p}}_i = \mathbf{K} \mathbf{p}_i^{\text{cam}}$, followed by perspective division:

$$\mathbf{p}_i^{\text{proj}} = \begin{pmatrix} \tilde{p}_{i,x} / \tilde{p}_{i,z} \\ \tilde{p}_{i,y} / \tilde{p}_{i,z} \end{pmatrix}, \quad (1)$$

where $\mathbf{p}_i^{\text{proj}} \in \mathbb{R}^2$ denotes the 2D pixel coordinate with image size (H, W) corresponding to the i -th 3D point. Subsequently, we extract local image evidence from the feature maps by mapping projected pixel coordinates to the corresponding feature map with $h \times w$ resolution. Rather than relying on a single projection-aligned feature that is sensitive to projection errors, we adopt bilinear grid sampling on the image feature map over a local neighborhood around each projected position, yielding $K * K$ (with $K=5$) candidate visual evidence features that capture local appearance variation.

After obtaining the local image features aligned with each 3D point, we aggregate the sampled features into a single point-level image feature using a geometry-aware weighted pooling. For the i -th 3D point, let $\{\mathbf{f}_{i,k} \in \mathbb{R}^C\}_{k=1}^K$ denote the set of local image features sampled from the feature map, with corresponding normalized coordinates $\{\mathbf{c}_{i,k} \in \mathbb{R}^2\}_{k=1}^K$. We compute a normalized 2D center for each 3D point $\bar{\mathbf{c}}_i$ from the sampled coordinates, and measure the relative distance between each sample and this center: $d_{i,k} = \|\mathbf{c}_{i,k} - \bar{\mathbf{c}}_i\|_2$. These distances are converted into spatial weights using a distance-based weighting function, assigning higher importance to the sampled features that are spatially closer to the projected point:

$$w_{i,k} = \frac{\exp(-d_{i,k}/\sigma)}{\sum_{k'} \exp(-d_{i,k'}/\sigma)}. \quad (2)$$

Therefore, the final point-aligned image feature is obtained as a weighted sum of the local image features: $\mathbf{f}_i^{\text{pa}} = \sum_{k=1}^K w_{i,k} \mathbf{f}_{i,k}$.

3.2.2 Learnable Tokenization via Encoding and Processing. Although geometry-aligned projection establishes explicit correspondence between image observations and 3D points, the resulting N_{point} point-aligned features $\mathbf{f}^{\text{pa}} \in \mathbb{R}^{N_{\text{point}} \times C}$ remain highly overcomplete. Due to redundant points and the limited spatial resolution of image features, many neighboring points receive nearly identical visual evidence. Consequently, directly treating point-aligned features as image tokens leads to representations dominated by redundant visual evidence rather than by informative visual content. To address this issue, we introduce a learnable tokenization module that compresses point-aligned features into a compact set of latent VETs. Specifically, we introduce a set of learnable query tokens $\mathbf{Q}_l$ to make tokenization a data-driven aggregation process. The learnable queries then enable the model to adaptively summarize the most informative and complementary point-aligned visual evidence into a fixed number of tokens through cross-attention:

$$\text{Attn}(\mathbf{Q}_l, \mathbf{f}^{\text{pa}}) = \text{softmax}\left(\mathbf{Q}_l \mathbf{f}^{\text{pa}^\top} / \sqrt{d}\right) \mathbf{f}^{\text{pa}}, \quad (3)$$

where $\mathbf{Q}_l \in \mathbb{R}^{N_{\text{vet}} \times C}$ serves as the query and $\mathbf{f}_{\text{pa}} \in \mathbb{R}^{N_{\text{point}} \times C}$ as the key and value. This encoding step adaptively aggregates the point-level features into N_{vet} token-level representations, effectively enabling the model to allocate representational capacity based on visual content. While this stage captures point-level evidence independently for each token, it does not explicitly promote coordination among tokens. To address this, we introduce a subsequent processing stage based on self-attention, allowing tokens to interact and exchange information. Finally, a feed-forward network is applied to further refine its representation, yielding a compact and content-adaptive set of VETs, noted as $\mathbf{T}_{\text{VET}}$, for subsequent modeling.

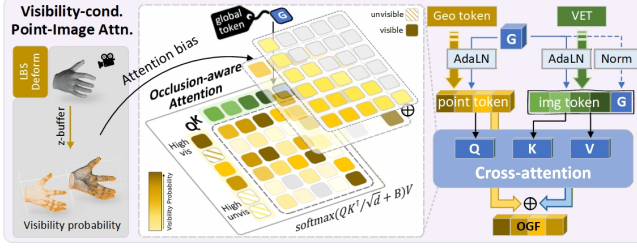


Figure 3: Architecture of the proposed Visibility-conditioned Point-Image Attention. OGF is short for occlusion-aware Gaussian features.

3.3 Occlusion-aware Gaussian Feature Construction

After constructing geometry-aligned VETs, our next goal is to reliably transfer their visual evidence to geometry-anchored point tokens for point-wise appearance inference. Prior avatar methods model 3D-2D interaction either through cross-attention [1, 2, 24] or unified multimodal transformers with full attention over geometric and visual tokens [27]. However, under severe hand self-occlusion, occluded points project onto visible foreground regions rather than their own surface. In this case, feature-affinity attention cannot determine whether the matched image evidence is reliable for the queried point, making mismatched point-image associations difficult to suppress and causing erroneous appearance attribution. To address the above issues, we introduce visibility-conditioned point-image attention (VPJA), which explicitly modulates how each 3D point attends to image evidence based on its estimated visibility. This design encourages visible points to focus on reliable local evidence, while allowing occluded points to rely more on global visual context. To explicitly model dependencies among Gaussian features, we apply self-attention to propagate contextual information across features, enabling more coherent and robust reconstruction.

3.3.1 Multimodal Token Representation. Alongside the VETs, we introduce point-based geometric tokens that provide the structural foundation for subsequent visibility-conditioned point-image interactions and joint reasoning. The geometric tokens are derived from MANO-HD surface points and encode structural priors of the hand. By initialization on the MANO-HD template, each point is embedded via sinusoidal positional encoding [16] followed by an multi-layer perceptron (MLP) projection: $\mathbf{T}_{3D} = \text{MLP}_{\text{proj}}(\gamma(\mathbf{p})) \in \mathbb{R}^{N_{\text{points}} \times C}$, where $\gamma: \mathbb{R}^3 \rightarrow \mathbb{R}^{3L}$ applies L -frequency sinusoidal encoding to spatial coordinates. To capture global context information, we take the global context features $\mathbf{g}$ extracted from DI-NOv2 as input, followed by max pooling and two MLP layers: $\mathbf{T}_{\text{global}} = \text{MLP}_{\text{global}}(\text{MaxPool}(\mathbf{g}))$. Following previous work [27, 37], we use global features to condition Adaptive Layer Normalization (AdaLN), producing separate modulation for geometric tokens and VETs during attention-based fusion.

3.3.2 Visibility-conditioned Point-Image Attention. With the geometric and visual tokens constructed, we next introduce a visibility-conditioned point-image attention module to regulate evidence transfer under self-occlusion. The key challenge is that the reliability of visual evidence is inherently visibility-dependent, as occluded points often lack reliable local image evidence. To address

this, we first estimate point visibility by a z-buffer test under the posed hand geometry, and complement the local VETs with a global appearance context to support occluded points. Conditioned on the estimated visibility and the local-global visual evidence, our visibility-conditioned point-image cross-attention emphasizes local cues for visible points while shifting attention toward global context for occluded points, yielding occlusion-aware visual context for Gaussian feature construction, as depicted in fig. 3.

Given a projected 3D point $\mathbf{p}_i^{\text{proj}}$ in the camera space with depth z_i , we sample the corresponding depth value $z_i^{\text{buf}}(\mathbf{p}_i^{\text{proj}})$ from the rendered depth map, where $z_i^{\text{buf}}(\cdot)$ denotes the depth-buffer value of the visible posed hand mesh at the queried image location. We then compute the depth difference and map it to a soft visibility score by sigmoid to obtain a continuous visibility score:

$$\Delta z_i = z_i^{\text{buf}}(\mathbf{p}_i^{\text{proj}}) - z_i \quad (4)$$

$$p_i^{\text{vis}} = \sigma(\Delta z_i / \tau)$$

where Δz_i is the signed depth residual between the point and the z-buffer depth at $\mathbf{p}_i^{\text{proj}}$. The sign encodes whether the point lies in front of or behind the visible surface, while the magnitude measures the depth deviation. τ controls the softness of the visibility transition. The resulting score $p_i^{\text{vis}} \in [0, 1]$ serves as a geometric proxy for the reliability of local image evidence associated with each 3D point.

To provide occluded points with robust appearance guidance, we introduce a global appearance context as an alternative source of image evidence. By augmenting the local VETs with a global visual token, they form a unified set of image tokens that jointly represent fine-grained local cues and coarse but robust global appearance information. Given this multi-source image evidence, the estimated visibility score naturally serves as a point-wise prior on the reliability of point-image correspondence that governs how each 3D point should select and weight the different sources of visual information. To this end, we incorporate visibility as an additive bias in the point-to-image cross-attention, allowing it to modulate relative attention preference. Concretely, given the soft visibility score p_i^{vis} for point i , we first map it to a signed scalar $s_i = 2p_i^{\text{vis}} - 1$, where $s_i \in [-1, 1]$ encodes the directional preference of visual evidence: positive values indicate reliable local cues, while negative values suggest local evidence is likely corrupted by occlusion and global context should be preferred. Based on this formulation, we define a visibility-conditioned attention bias $b_{i,j}$ between point i and image token j as:

$$b_{i,j} = \begin{cases} \beta_{\text{local}} s_i, & j \in \mathcal{J}_{\text{VET}} \\ -\beta_{\text{global}} s_i, & j \in \mathcal{J}_{\text{global}} \end{cases} \quad (5)$$

where $\mathcal{J}_{\text{VET}}$ and $\mathcal{J}_{\text{global}}$ denote the index sets of local VETs and the global token, and β_{local} and β_{global} control the bias strength for local and global tokens, respectively. The opposite signs encode a visibility-conditioned preference over local and global image evidence. The visibility-conditioned bias is then added to the attention logits in the point-to-image cross-attention:

$$\tilde{\mathbf{T}}_{3D} = \text{softmax}(\mathbf{Q}\mathbf{K}^T / \sqrt{d} + \mathbf{B}) \mathbf{V} \quad (6)$$

$$\mathbf{Q} = W_q \mathbf{T}_{3D}, \quad \tilde{\mathbf{T}}_{\text{VET}} = \mathbf{T}_{\text{VET}} + \mathbf{P}_{\text{VET}}, \quad [\mathbf{K}, \mathbf{V}] = W_{kv} \mathbf{T}_{2D}^{\text{full}}.$$

Here, geometric tokens $\mathbf{T}_{3D}$ serve as the query, and $\mathbf{T}_{2D}^{\text{full}} = [\tilde{\mathbf{T}}_{\text{VET}}; \mathbf{T}_{\text{global}}]$ are the key and value. W_q and W_{kv} are learnable projections that

map the $\mathbf{T}_{3D}$ to the query $\mathbf{Q}$ and the visual tokens to the key $\mathbf{K}$ and value $\mathbf{V}$, respectively, and d is the dimension of the query and key vectors. $\mathbf{P}_{\text{VET}}$ encodes the 2D location of VETs. Therefore, $\mathbf{Q}\mathbf{K}^\top$ reflects the spatial correspondence between each 3D geometric query and the 2D visual tokens, indicating how relevant each visual token is to the queried point. After softmax normalization, these correspondence weights are used to aggregate the value features, allowing each queried point to read visual evidence from the attended visual tokens. $\mathbf{B}$ is the bias matrix formed by $b_{i,j}$, and $\tilde{\mathbf{T}}_{3D}$ is the occlusion-aware visual information. Consequently, the resulting point-to-image attention weights allocate more mass to local tokens for visible points ($s_i > 0$) and shift mass towards the global token for occluded points ($s_i < 0$). By injecting visibility-conditioned bias at the logit level, our method effectively modulates evidence reading, encouraging visible points to attend to local visual tokens while allowing occluded points to rely more on global context. The resulting occlusion-aware visual context are then added to the original geometric tokens to form the occlusion-aware Gaussian features.

3.3.3 Self-Attention Refinement of Occlusion-Aware Gaussian Features. Although visibility-conditioned cross-attention injects image evidence into geometry tokens, it updates each token independently and therefore cannot explicitly capture dependencies among the Gaussian features. Since the resulting occlusion-aware Gaussian features encode both geometric and occlusion-aware visual context, these two types of information should be further propagated and coordinated across features for coherent reconstruction. Therefore, we apply self-attention to the occlusion-aware Gaussian features, allowing each Gaussian to be exchanged and refined with structural and visual context from others. This inter-feature propagation improves the contextual consistency of Gaussian representations.

3.4 Feature-on-Mesh Representation

Building on the occlusion-aware Gaussian features, we further introduce a more suitable representation carrier for Gaussian deformation under articulated hand animation. Existing approaches [27, 35] typically rely on vertex-based features for Gaussian decoding, implicitly assuming that vertex-level representations are sufficient to capture both appearance and deformation. However, for highly articulated hands, non-rigid deformation often manifests as local stretching and shearing over mesh faces rather than only at vertices. As a result, decoding Gaussians solely from vertex-wise features may fail to adequately reflect local surface deformation over mesh faces, limiting pose-dependent appearance fidelity. To address this limitation, we introduce a Feature-on-Mesh (FoM) representation as a face-level complement to the fused features.

For each mesh triangle t_i with vertices $\{v_{i,1}, v_{i,2}, v_{i,3}\}$, we introduce a set of learnable barycentric coordinate (u_i^F, v_i^F) , normalized through a softmax to ensure a valid convex combination, to interpolate vertex-wise features within each face, yielding a face-embedded representation that better captures local surface variation than fixed face sampling. Concretely, given the vertex-wise latent features $\{\mathbf{F}_{i,1}, \mathbf{F}_{i,2}, \mathbf{F}_{i,3}\}$ obtained from the occlusion-aware Gaussian features, we compute the face-level feature as a barycentric interpolation of the corresponding vertex features:

$$\mathbf{F}_i^{\text{FoM}} = u_i^F * \mathbf{F}_{i,1} + v_i^F * \mathbf{F}_{i,2} + (1 - u_i^F - v_i^F) * \mathbf{F}_{i,3}. \quad (7)$$

Each face-level feature is then decoded into Gaussian attributes, including opacity, color, scaling, rotation, and a set of barycentric coordinates (u_i^G, v_i^G) for Gaussian positioning within each triangle. In this way, the decoded Gaussians are explicitly tied to the mesh surface. During animation, we first deform the MANO-HD mesh with LBS and then update Gaussian positions by barycentric interpolation of the posed triangle vertices, ensuring surface-consistent motion. To further adapt Gaussian shape to local non-rigid deformation, we adopt a mesh-driven deformation scheme inspired by [31], where face-level rotations are derived from mesh deformation and used to guide Gaussian transformation (more details are shown in appendix B). By representing features on mesh faces and coupling Gaussian deformation to mesh deformation, FoM provides a more suitable carrier for pose-dependent Gaussian modeling, leading to consistent appearance reconstruction under hand articulation.

After obtaining the point-wise features and mesh-embedded features, we employ a Gaussian Head to predict 3DGS parameters. For points-wise features, we have:

$$\begin{aligned} \{\Delta \mathbf{x}_i, \mathbf{r}_i, \mathbf{c}_i, \mathbf{s}_i, \sigma_i\} &= \text{MLP}_{\text{GS}}(\mathbf{T}_{3D}^i) \\ \mathbf{p}_i &= \mathbf{p}_i + \Delta \mathbf{x}_i, \quad \forall i \in \{1, \dots, N_{\text{points}}\} \end{aligned} \quad (8)$$

where $\Delta \mathbf{x}_i \in \mathbb{R}^3$ represents residual position offsets from the canonical MANO-HD. For mesh-embedded features, we have:

$$\begin{aligned} \{u_j, v_j, \mathbf{r}_j, \mathbf{c}_j, \mathbf{s}_j, \sigma_j\} &= \text{MLP}_{\text{FoM}}(\mathbf{F}_j^{\text{FoM}}) \quad \forall j \in \{1, \dots, N_{\text{FoM}}\} \\ \mathbf{p}_j &= u_j * \mathbf{p}_{j,1} + v_j * \mathbf{p}_{j,2} + (1 - u_j - v_j) * \mathbf{p}_{j,3}. \end{aligned} \quad (9)$$

Training and one-shot strategies are shown in appendix B.

4 Experiments

4.1 Implementation Details

Learning OASIS. We adopt 21 subjects from the InterHand2.6M [18] training set for pretraining, following [41]. During training, we randomly sample a source view image and four target view images from a subject. Our network is trained on two NVIDIA A100 40GB using the Adam optimizer [14] with the learning rates of 4×10^{-4} . Loss weights for training objective are set as $\lambda_1 = 10, \lambda_2 = \lambda_3 = 1, \lambda_4 = 5$ and $\lambda_p = 1, \lambda_s = 0.5$.

One-shot Reconstruction. We conduct one-shot reconstruction evaluations on the testing set of InterHand2.6M as in [3, 41] for a fair comparison. For in-the-wild images, we utilize an off-the-shelf estimator [23] to predict the MANO parameters and camera poses. We set $\lambda_\mu = 20, \lambda_\sigma = 1$. The one-shot learning takes 100 steps for color calibration and 400 steps for texture details learning. More details about the experiments are shown in appendix B.

4.2 Evaluation of One-shot Reconstruction

Quantitative Comparison. We quantitatively compare OASIS with previous methods [3, 10, 11, 19, 22, 34, 41]. Table 1 summarizes the quantitative results of our model against the baselines on the testing set of InterHand2.6M [18]. We additionally include a no-finetuning variant of our method (denoted as $\S$), which directly predicts 3D Gaussians from a single input image without one-shot adaptation, to quantify the contribution of the one-shot learning and assess the strength of the prior net. To further contextualize the benefit of our proposed framework, we compare our no-finetuning results with [27], where we train it on the same setting as our prior

Table 1: Quantitative results on InterHand2.6M. We highlight the best and second-best results of each metric. LPIPS* = $100 \times \text{LPIPS}$. § means testing without finetuning.

	Method	#Train	PSNR↑	SSIM↑	LPIPS*↓
Multi-img	SelfRecon [10]	11,757	26.38	0.879	14.21
	HumanNeRF [34]	11,757	27.64	0.884	11.45
	HandAvatar [3]	11,757	28.23	0.894	10.35
Single-img	HARP [11]	1	19.82	0.761	22.49
	Handy [22]	1	25.56	0.794	14.98
	HandAvatar [3]	1	23.79	0.820	17.78
	OHTA [41]	1	26.11	0.864	12.93
	LHM§ [27]	-	24.45	0.901	13.96
	Ours§	-	26.48	0.950	11.74
	Ours	1	27.38	0.956	11.45

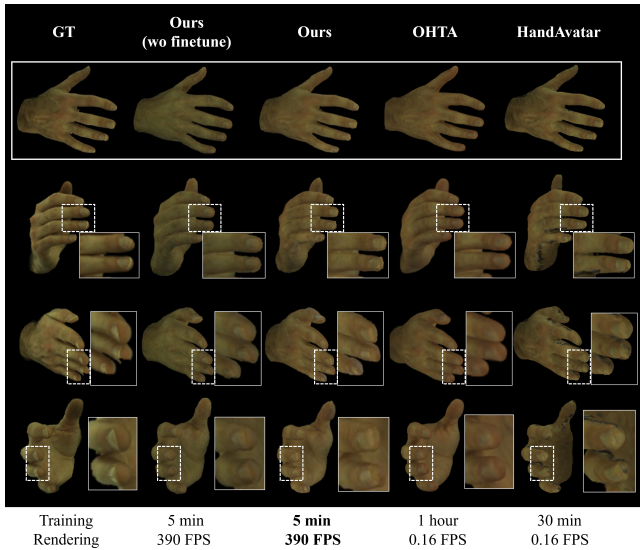


Figure 4: Qualitative comparison with state-of-the-art methods on InterHand2.6M [18]. The white box indicates the input image. Frames per second (FPS) are measured on an NVIDIA A100 GPU.

net training for a fair comparison. Overall, our one-shot reconstruction achieves the superior performance among one-shot methods across all metrics, while the no-finetuning variant still surpasses the baselines, validating the effectiveness of the proposed prior net. Although methods trained with monocular videos [3, 10, 34] benefit from substantially stronger input supervision, our one-shot results are the most competitive with them, further demonstrating strong performance under single-image observations. More quantitative comparisons on HanCo dataset [42] are shown in appendix C.3.

Qualitative Comparison. As shown in fig. 4, our OASIS demonstrates superior reconstruction quality in InterHand2.6M [18] compared to baseline models. Although HandAvatar [3] shows impressive results with monocular videos, it fails to generalize well in one-shot scenarios due to the information missing from the single input image. With the prior net and one-shot reconstruction

pipeline, OHTA [41] achieves comparable visualization results to our OASIS for $10\times$ more time to converge (around 1 hour). However, it does not perform well on high-frequency texture details, particularly evident in nails. Through our proposed 3D Gaussian-based representation framework, we achieve high-fidelity reconstruction with rapid one-shot learning in approximately 5 minutes. Additionally, our method enables fast rendering at 390 FPS, whereas [3, 41] require roughly $2500\times$ more time to render a single image (both operating at around 0.16 FPS). Moreover, our model without finetuning exhibits color discrepancies due to the unseen subject during training, which further validates the effectiveness of the one-shot reconstruction pipeline. Additional qualitative comparisons on the HanCo dataset [42] are provided in appendix C.4.

4.3 Ablation study

Tab. 2 summarizes the quantitative results of the effectiveness for each component, which are performed on Interhand2.6M dataset. For each ablated variant, we first train the corresponding model with the target component removed, then perform one-shot personalization using the resulting model, and finally evaluate it under the same one-shot setting. Fig. 5 presents qualitative results to highlight their contributions. More ablation studies on one-shot reconstruction strategy are shown in appendix C.6.

w/o VET. To validate the effectiveness of the proposed geometry-aligned visual evidence tokens construction, we conduct ablation study #b, which removes the VETs and directly applies image features from the backbone instead, as suggested in [27]. Fig. 5 b shows that removing VETs leads to blurry local structures and weak fine-grained appearance recovery of fine-grained appearance details, with noticeable distortions in high-frequency regions such as the fingertips and fingernails. In contrast, our full model preserves clearer contours and richer local detail by explicitly aligning image evidence with the corresponding 3D hand geometry through geometry-aligned VETs, which reduces ambiguity in associating observed appearance cues with the correct 3D geometry. This visual improvement is consistent with the better PSNR and LPIPS reported in tab. 2 #b. Additionally, we conduct two further ablation studies on the aggregation and LTEP modules. Specifically, we remove the aggregation step and directly use the projected features to assess the importance of feature aggregation. For the LTEP module, we replace it with a MLP to evaluate its contribution. In both cases, tab. 2 #b.1 and #b.2 verify that removing the corresponding module leads to degraded performance. The former improves the consolidation of point-aligned local visual cues into more reliable geometry-aligned evidence, while the latter converts the resulting dense evidence into compact and informative visual tokens for subsequent fusion. The performance drop in both cases confirms that both components are necessary for the effectiveness of the complete VETs design.

w/o VPIA. For robust image evidence transfer under severe self-occlusion, we introduce a visibility-conditioned point-image attention. To evaluate its benefit, we perform an ablation study #c to remove VPIA module and instead apply a multimodal transformer in [27] to jointly model geometric tokens and VETs. In addition, we further conduct a bias-only ablation, where the visibility bias in VPIA is removed while keeping the remaining attention formulation unchanged. Without the full VPIA mechanism, the reconstruction

#	Method	PSNR↑	LPIPS*↓	SSIM↑
a	Full Model	27.38	11.45	0.956
b	w/o VETs	26.97	12.11	0.955
b.1	w/o Aggr.	27.08	11.87	0.955
b.2	w/o LTEP	27.12	11.98	0.955
c	w/o VPIA	27.10	12.09	0.955
c.1	w/o bias	27.07	12.08	0.955
d	w/o FoM	27.13	12.35	0.955

Table 2: Ablation studies of OASIS under the one-shot setting.

becomes less reliable in occluded regions, where the partially occluded fingernail area exhibits locally inconsistent structure and texture, as shown in fig. 5 c. Removing only the bias leads to noticeable degradation in fig. 5 c.1, where chaotic textures appear in occluded fingertip regions, indicating that the visibility-conditioned bias is crucial for suppressing unreliable local evidence for occluded regions. Our full model alleviates this issue by modulating image evidence transfer according to visibility, allowing visible points to incorporate reliable local image cues while encouraging occluded points to rely more on global information. This advantage is further supported by the better PSNR and LPIPS in tab. 2 #c and #c.1.

w/o FoM. To better support hand deformation, we introduce the FoM representation, which lifts features onto mesh faces to provide a face-level basis for Gaussian modeling. To prove the contribution of the FoM, we ablate this representation in experiment #d, which only utilizes the fused point features for decoding and animation. As shown in fig. 5 d, removing FoM causes noticeable artifacts during pose changes, particularly local holes in regions undergoing surface stretching, where the Gaussian representation fails to maintain sufficient surface coverage. By contrast, our full model preserves continuous hand surfaces and structurally stable local regions by using FoM as a face-level feature carrier that drives the decoded Gaussians to deform consistently with mesh face deformation, yielding coherent appearance across poses without noticeable holes or discontinuities. This visual difference is consistent with the significant drop in LPIPS in tab. 2 #d, suggesting that removing FoM representation weakens the model’s ability to preserve a coherent and stable appearance during non-rigid deformation.

4.4 Robustness towards Diverse Input and Applications

To demonstrate the robustness of OASIS, we conduct experiments on diverse input images both quantitatively and qualitatively. The quantitative comparisons are shown in appendix C.5. Furthermore, we show more practical applications on text-to-avatar, where the hand avatar is reconstructed by the text prompts, and hand avatar editing. To illustrate the effectiveness of our model, we make comprehensive comparisons with [41] on every input image and animation result. In appendix fig. 6, we present single-image 3D hand reconstructions on various in-the-wild images, where the first row for each image shows our reconstruction avatars and the second row shows the results reconstructed from OHTA [41]. Notably, OHTA struggles under highly articulated hand poses (e.g., the first

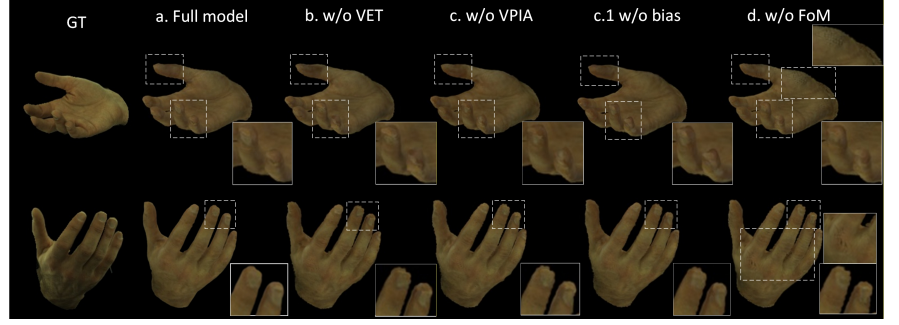


Figure 5: Visual results for ablation studies of OASIS on InterHand2.6M dataset.

row), indicating limited robustness to large pose variations. In contrast, our method preserves richer and more faithful fine-grained appearance details, including palm wrinkles and ring structures. Additionally, we animate the reconstructed avatar for comparison, where OHTA tends to produce texture artifacts along the boundary between observed and unobserved regions, whereas our method maintains consistent textures and smooth transitions. These results demonstrate that our approach generalizes effectively to unseen poses, yielding high-fidelity textures and coherent animation.

Based on our one-shot reconstruction pipeline, we further demonstrate two downstream applications: text-to-avatar and hand avatar editing. For hand avatar editing, we can draw arbitrary content on any single image and then utilize OASIS to reconstruct the target hand avatar, as illustrated in fig. 1. For text-to-avatar, we generate hand images through a hand mask and text prompts following [41], and then reconstruct animatable 3D hand avatars from the generated images. The resulting avatars support user-specified animation while preserving the input-view appearance. As shown in fig. 7 in appendix, [41] tends to smooth out textures under such synthesized images and often fails to reconstruct high-frequency details (e.g., accessories and nails). In comparison, our method preserves finer-grained textures in the input view and maintains consistent appearance during animation, resulting in improved identity and texture fidelity. These results demonstrate that OASIS is more robust to diverse input appearances and can maintain better faithful reconstruction with animation consistency in one-shot settings.

5 Conclusion

In this work, we propose a hand-tailored 3DGS model for reconstructing animatable hand avatars from a single image. To address the sparse appearance evidence and severe self-occlusion, we introduce geometry-aligned visual evidence tokens to preserve image-specific appearance cues and a visibility-conditioned point-image attention module to extract occlusion-aware visual context. By combining this visual context with geometric tokens, we construct occlusion-aware Gaussian features, enabling faithful and robust reconstruction under challenging poses and self-occlusion. Furthermore, our Feature-on-Mesh representation enables Gaussian deformation to better follow local surface deformation, improving reconstruction consistency. Extensive experiments demonstrate that OASIS achieves superior reconstruction quality and efficiency compared with existing approaches, highlighting its effectiveness for animatable hand avatar reconstruction from a single image.

References

- [1] Jinnan Chen, Chen Li, Jianfeng Zhang, Lingting Zhu, Buzhen Huang, Hanlin Chen, and Gim Hee Lee. 2024. Generalizable Human Gaussians from Single-View Images. *arXiv preprint arXiv:2406.06050* (2024).
- [2] Wenyue Chen, Peng Li, Wangguandong Zheng, Chengfeng Zhao, Mengfei Li, Yaolong Zhu, Zhiyang Dou, Ronggang Wang, and Yuan Liu. 2025. SyncHuman: Synchronizing 2D and 3D Generative Models for Single-view Human Reconstruction. *arXiv preprint arXiv:2510.07723* (2025).
- [3] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. 2023. HandAvatar: Free-Pose Hand Animation and Rendering from Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8683–8693.
- [4] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. 2022. Lisa: Learning implicit shape and appearance of hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20533–20543.
- [5] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando D De la Torre. 2024. Hamba: Single-view 3D Hand Reconstruction with Graph-guided Bi-Scanning Mamba. *Advances in Neural Information Processing Systems* 37 (2024), 2127–2160.
- [6] Zhiyang Guo, Wengang Zhou, Min Wang, Li Li, and Houqiang Li. 2023. Hand-NeRF: Neural Radiance Fields for Animatable Interacting Hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21078–21087.
- [7] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. 2025. LAM: Large Avatar Model for One-shot Animatable Gaussian Head. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–13.
- [8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr* 1, 2 (2022), 3.
- [9] Shoukang Hu, Tao Hu, and Ziwei Liu. 2024. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20418–20431.
- [10] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. 2022. SelfRecon: Self Reconstruction of Your Digital Avatar from Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5605–5615.
- [11] Korrawe Kar, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. 2023. HARP: Personalized Hand Reconstruction from a Monocular RGB Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12802–12813.
- [12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.
- [13] Minje Kim and Tae-Kyun Kim. 2024. Bitt: Bi-directional texture reconstruction of interacting two hands from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10726–10735.
- [14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [15] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. 2022. Nimble: a non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–16.
- [16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.
- [17] Gyeongsik Moon, Weipeng Xu, Rohan Joshi, Chenglei Wu, and Takaaki Shiratori. 2024. Authentic Hand Avatar from a Phone Scan via Universal Hand Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2029–2038.
- [18] Gyeongsik Moon, Shoo-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*. Springer, 548–564.
- [19] Akshay Mundra, Jiayi Wang, Marc Habermann, Christian Theobalt, and Mohamed Elgharib. 2023. LiveHand: Real-Time and Photorealistic Neural Hand Rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 18035–18045.
- [20] Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, and Minh Hoai. 2019. Contextual attention for hand detection in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9567–9576.
- [21] Maxime Quab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2024. DiNov2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal* (2024).
- [22] Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, and Stefanos Zafeiriou. 2023. Handy: Towards a High-Fidelity 3D Hand Shape and Appearance Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4670–4680.
- [23] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. 2025. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 12242–12254.
- [24] Lorenza Prospero, Abdullah Hamdi, Joao F Henriques, and Christian Rupprecht. 2025. Gst: Precise 3d human body from a single image with gaussian splatting transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 6007–6017.
- [25] Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. 2020. Html: A parametric hand texture model for 3d hand reconstruction and personalization. In *European Conference on Computer Vision*. Springer, 54–71.
- [26] Zhiyin Qian, Shaoifei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 2024. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5020–5030.
- [27] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. 2025. LHM: Large Animatable Human Reconstruction Model for Single Image to 3D in Seconds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14184–14194.
- [28] Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, et al. 2025. Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 21148–21158.
- [29] Javier Romero, Dimitrios Tzionas, and Michael J Black. 2017. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 1–17.
- [30] Umme Sara, Morium Akter, Mohammad Shorif Uddin, et al. 2019. Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications* 7, 3 (2019), 8–18.
- [31] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. 2024. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1606–1616.
- [32] Zhigang Tu, Zhisheng Huang, Yujin Chen, Di Kang, Linchao Bao, Bisheng Yang, and Junsong Yuan. 2023. Consistent 3d hand reconstruction in video via self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 8 (2023), 9469–9485.
- [33] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
- [34] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. 2022. HumanNeRF: Free-Viewpoint Rendering of Moving People from Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16210–16220.
- [35] Yufan Wu, Xuanhong Chen, Wen Li, Shunran Jia, Hualiang Wei, Kairui Feng, Jialiang Chen, Yuhao Li, Ang He, Weimin Zhang, et al. 2025. SinGS: Animatable Single-Image Human Gaussian Splats with Kinematic Priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5571–5580.
- [36] Yue Wu, Yufan Wu, Wen Li, Yuxi Lu, Kairui Feng, and Xuanhong Chen. 2025. FastAvatar: Towards Unified Fast High-Fidelity 3D Avatar Reconstruction with Large Gaussian Reconstruction Transformers. *arXiv preprint arXiv:2508.19754* (2025).
- [37] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).
- [38] Dongbin Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Yang Li, Minghan Qin, Yu Li, and Haoqian Wang. 2025. GUAVA: Generalizable Upper Body 3D Gaussian Avatar. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14205–14217.
- [39] Ruiqi Zhang and Jie Chen. 2024. Mesh-centric gaussian splatting for human avatar modelling with real-time dynamic mesh reconstruction. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 6823–6832.
- [40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.
- [41] Xiaozheng Zheng, Chao Wen, Zhuo Su, Zeran Xu, Zhaohu Li, Yang Zhao, and Zhou Xue. 2024. OHTA: One-Shot Hand Avatar via Data-Driven Implicit Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 799–810.
- [42] Christian Zimmermann, Max Argus, and Thomas Brox. 2021. Contrastive representation learning for hand shape estimation. In *DAGM German Conference on Pattern Recognition*. Springer, 250–264.

Appendix

A Preliminary

3D Gaussian Splatting [12] is an explicit 3D representation based on a set of Gaussian primitives, which supports real-time rendering through differentiable rasterization. Specifically, each Gaussian primitive is parameterized by its position center μ and a full 3D covariance matrix Σ in a world space, i.e.,

$$G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)}, \quad (10)$$

where $\mathbf{x}$ is any position in 3D space. To guarantee that the covariance matrix Σ is positive semi-definite, it is factorized into a scaling matrix $\mathbf{S}$ and a rotation matrix $\mathbf{R}$, defined as $\Sigma = \mathbf{R}\mathbf{S}\mathbf{S}^\top\mathbf{R}^\top$. In implementation, each Gaussian is described by a diagonal scaling vector $\mathbf{s} \in \mathbb{R}^3$ and a quaternion vector $\mathbf{r} \in \mathbb{R}^4$, from which a valid covariance matrix can be readily reconstructed. Furthermore, each Gaussian is associated with an opacity value $\alpha \in [0, 1]$ and a color vector $\mathbf{c} \in \mathbb{R}^3$. As a result, a set of 3D Gaussians is represented by the parameters $\mathcal{G} = \{\mu, \mathbf{s}, \mathbf{r}, \mathbf{c}, \alpha\}$. After projecting 3D Gaussians onto the 2D image plane, denoted as G' , the color of a pixel p is obtained by α -blending the N ordered Gaussians that overlap with it:

$$\hat{I}(p) = \sum_{i=1}^N c_i \sigma_i \prod_{j=1}^{i-1} (1 - \sigma_j), \quad \sigma_i = \alpha_i G'_i(p), \quad (11)$$

where σ_i is the opacity contribution of the i -th Gaussian on the pixel p .

MANO [29] is a pretrained parametric hand model representing hand shape and pose by defining θ, β . It has been widely adopted in prior hand modeling and avatar works [3, 4, 11, 41] to model articulated hand deformation. In this work, we use MANO to animate canonical 3D Gaussians from canonical space to posed space. Specifically, the 3D position and covariance of each 3D Gaussian is translated and rotated by the estimated LBS transformation matrix,

$$\begin{aligned} \mathbf{p}^t &= \mathbf{G}(\mathbf{J}^t, \theta^t) \mathbf{p}^c + \mathbf{b}(\mathbf{J}^t, \theta^t, \beta^t) \\ \Sigma^t &= \mathbf{G}(\mathbf{J}^t, \theta^t) \Sigma^c \mathbf{G}(\mathbf{J}^t, \theta^t)^\top, \end{aligned} \quad (12)$$

where $\mathbf{p}^t$, Σ^t , $\mathbf{p}^c$, and Σ^c are the position vector and covariance matrix in the posed and canonical space, respectively. $\mathbf{G}(\mathbf{J}^t, \theta^t) = \sum_{k=1}^K w_k \mathbf{G}_k(\mathbf{J}^t, \theta^t)$, $\mathbf{b}(\mathbf{J}^t, \theta^t, \beta^t) = \sum_{k=1}^K w_k \mathbf{b}_k(\mathbf{J}^t, \theta^t, \beta^t)$ are the rotation matrix and translation vector. K is the joint number, $\mathbf{G}_k(\cdot)$ and $\mathbf{b}_k(\cdot)$ are the transformation matrix and translation vector of joint k respectively, w_k is the LBS weight.

B Implementation details of OASIS

B.1 FoM Representation.

The FoM representation organizes features on mesh faces, serving as a face-level bridge between the Gaussian representation and local mesh deformation. However, during animation, simply posing Gaussians with LBS is insufficient to capture the local non-rigid deformation induced by hand articulation. To bridge this gap, we further use mesh-derived deformation to adapt Gaussian shapes according to local mesh stretching. Specifically, we adopt a mesh-embedded Gaussian deformation scheme to estimate per-vertex local rotations by area-weighted averaging over adjacent triangles, and interpolate them within each face to obtain a face-level rotation.

This face-level rotation is composed with the decoded canonical Gaussian rotation $\mathbf{q}_i^{\text{cano}}$ to determine the posed Gaussian orientation $\mathbf{q}_i^{\text{posed}}$:

$$\begin{aligned} \mathbf{q}_v &= \frac{\sum_{k \in \Omega_v} A_k \mathbf{q}_k}{\sum_{k \in \Omega_v} A_k} \\ \delta \mathbf{q}_{i,t} &= u_i^G \mathbf{q}_{v_{i,1}} + v_i^G \mathbf{q}_{v_{i,2}} + (1 - u_i^G - v_i^G) \mathbf{q}_{v_{i,3}} \\ \mathbf{q}_i^{\text{posed}} &= \delta \mathbf{q}_{i,t} * \mathbf{q}_i^{\text{cano}} \end{aligned} \quad (13)$$

where Ω_v is the neighbor triangles of vertex v , A_k and $\mathbf{q}_k$ are the triangle's area and quaternion respectively. Similarly, Gaussian scaling is modulated by the relative area change of each embedded triangle between canonical and posed space: $\mathbf{s}_i^{\text{posed}} = (A_i^{\text{posed}}/A_i^{\text{cano}}) \cdot \mathbf{s}_i^{\text{cano}}$, where A_i^{cano} and A_i^{posed} denote the triangle area in canonical and posed space, respectively.

B.2 Training Strategy

Given the predicted 3DGS attributes $\mathcal{G} = \{\mathbf{p}, \mathbf{r}, \mathbf{c}, \mathbf{s}, \sigma\}$, representing Gaussian position, rotation, color, scaling, and opacity, respectively, the canonical hand Gaussians are first transformed to target poses by LBS deformation. We then render the deformed Gaussians by differentiable splatting to obtain the rendered image $\hat{I}$ and the rendered mask $\hat{M}$. The view-consistent supervision comprises four components in view space:

$$\mathcal{L}_{\text{rec}} = \lambda_1 \mathcal{L}_{\text{rgb}} + \lambda_2 \mathcal{L}_{\text{mask}} + \lambda_3 \mathcal{L}_{\text{ssim}} + \lambda_4 \mathcal{L}_{\text{lpips}}. \quad (14)$$

RGB Loss. We use an $\mathcal{L}_1$ loss to compute pixel-wise error between the ground truth image I_{gt} and the rendered image $\hat{I}$:

$$\mathcal{L}_{\text{rgb}} = \|\hat{I} - I_{gt}\|_2. \quad (15)$$

Mask Loss. We apply the mask loss between the ground truth mask M and the accumulated volume density $\hat{M}$, defined as:

$$\mathcal{L}_{\text{mask}} = \|\hat{M} - M\|_2, \quad (16)$$

SSIM Loss. We also use SSIM to ensure the structural similarity between the rendered image and the ground truth image:

$$\mathcal{L}_{\text{ssim}} = \text{SSIM}(\hat{I}, I_{gt}), \quad (17)$$

LPIPS Loss. To ensure the quality of the rendered image, we utilize the LPIPS perceptual loss, i.e.,

$$\mathcal{L}_{\text{lpips}} = \text{LPIPS}(\hat{I}, I_{gt}), \quad (18)$$

In addition to the reconstruction loss, we introduce additional regularization terms to enforce geometric coherence in canonical space, as suggested in [27]. To maintain hand surface plausibility, we encourage Gaussian positions to be close to their initialized locations:

$$\mathcal{L}_{\text{position}} = \frac{1}{N_{\text{points}}} \sum_{i=1}^{N_{\text{points}}} \max(\|\Delta p_i\|_2 - d, 0) \quad (19)$$

where d is an empirically determined threshold. To penalize excessive thin and long Gaussian primitives, we apply a scaling regularization:

$$\ell_i = \max\left(\frac{s_i^{\text{max}}}{s_i^{\text{min}}} - r, 0\right), \quad \mathcal{L}_{\text{scaling}} = \frac{1}{N_{\text{point}}} \sum_{i=1}^{N_{\text{point}}} \ell_i \quad (20)$$

where $s_i \in \mathbb{R}^3$ is the Gaussian's scaling, and $s_i^{\text{max}}, s_i^{\text{min}}$ are the maximum and minimum scaling values respectively. r is an empirically

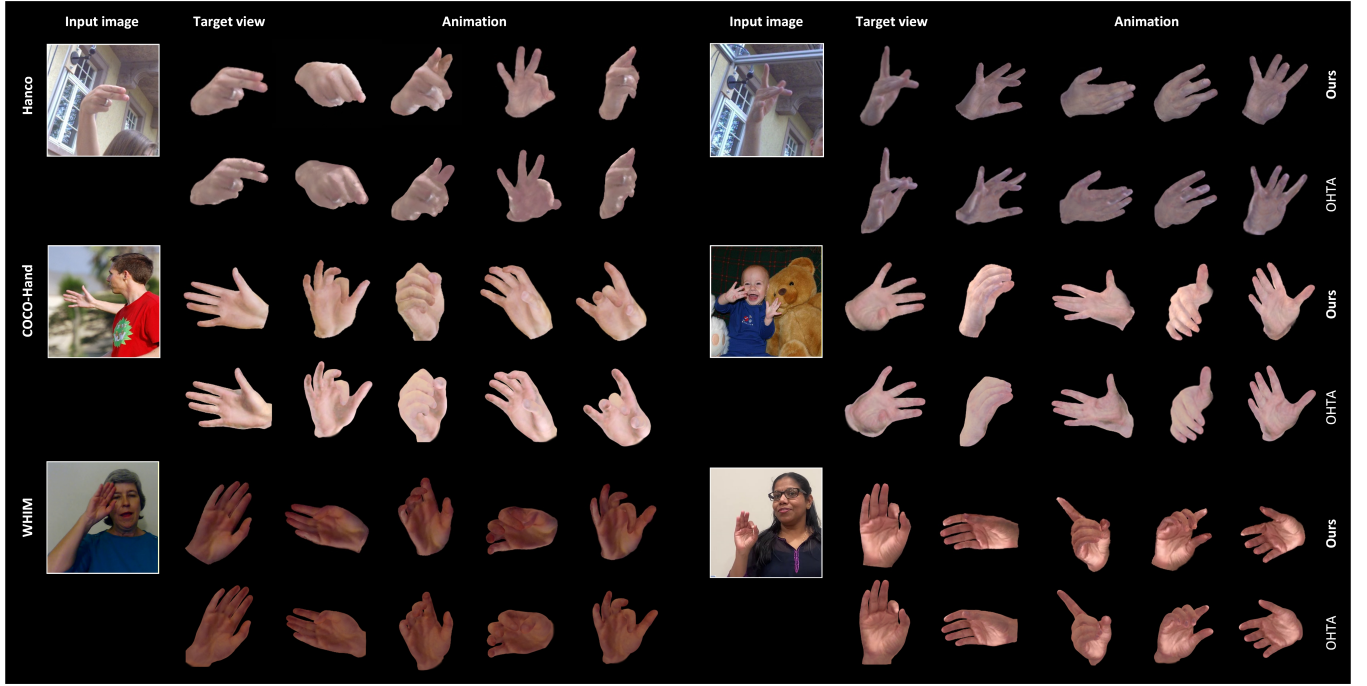


Figure 6: In the wild visualization from the HanCo [42] dataset, COCO-Hand [20] dataset, and WHIM [23] dataset.

determined threshold. Therefore, the total loss for training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_p \mathcal{L}_{\text{position}} + \lambda_s \mathcal{L}_{\text{scaling}}. \quad (21)$$

B.3 One-shot Hand Avatar Reconstruction

In the stage of one-shot hand avatar reconstruction, we first utilize an off-the-shelf hand pose estimator [23] to get the predicted hand shape parameter β , pose parameter θ and camera pose. Inspired by previous work [41], we optimize to render an image $\hat{I}$ that is similar to the target subject of the input image by introducing per-channel color calibration coefficients. Specifically, we introduce $\{\mathbf{w}, \mathbf{b}\} \in \mathbb{R}^{1 \times 3}$ to modulate the Gaussian appearance by $\tilde{\mathbf{c}} = \mathbf{w} * \mathbf{c} + \mathbf{b}$. We only optimize these color-related global parameters under masked photometric and color-statistics supervision, while freezing geometry and texture-related modules to preserve the pretrained appearance prior. For supervision, we compute the masked mean and standard deviation of an image, and enforce the rendered image to match the ground-truth image by minimizing the difference between their masked statistics:

$$\mu(I; M) = \frac{\sum I \odot M}{\sum M} \quad \sigma(I; M) = \sqrt{\frac{\sum (I - \mu(I; M))^2 \odot M}{\sum M}} + \epsilon$$

$$\mathcal{L}_{\text{calib}} = \lambda_1 \mathcal{L}_{\text{rgb}} + \lambda_\mu \left\| \mu(\hat{I}) - \mu(I_{gt}) \right\|_1 + \lambda_\sigma \left\| \sigma(\hat{I}) - \sigma(I_{gt}) \right\|_1 \quad (22)$$

After the color calibration stage, we train a LoRA (Low-Rank Adaptation) [8] module for Gaussian Head to capture the details of the target identity from the input image while keeping others frozen. To avoid overfitting the target view, we also perform view regularization. Specifically, we constrain the finetuning results of some

reference views with different poses $\{\mathbf{R}_i\}_{i=1}^{N^r}$ to be close to the rendering results with the same poses at the end of the color calibration stage.

$$\mathcal{L}_{\text{img}} = \mathcal{L}_{\text{rec}}(I, \hat{I}) + \lambda_{\text{ref}} \sum_{i=1}^{N^r} \left(\mathcal{L}_{\text{rec}}(I_i(\mathbf{R}_i), \hat{I}(\mathbf{R}_i)) \right) \quad (23)$$

where N^r is the number of the generated reference views. To prevent Gaussians from excessive anisotropy, we also perform the scaling regularization in Eq. 20.

$$\mathcal{L}_{\text{tex}} = \mathcal{L}_{\text{img}} + \lambda_s \mathcal{L}_{\text{scaling}} \quad (24)$$

C Experiments

C.1 Metrics

Following prior works [3, 41], we use Learned Perceptual Similarity (LPIPS) [40], Structural Similarity Index Measure (SSIM) [33], and Peak Signal-to-Noise Ratio (PSNR) [30] as reconstruction quality metrics to measure image similarity between ground truth images and rendered images.

C.2 Dataset Details

InterHand2.6M [18]. For the full model pretraining, we follow [41] to use the subject ‘train/Capture0’, ‘train/Capture1’, ‘train/Capture2’, ‘train/Capture3’, ‘train/Capture5’, ‘train/Capture6’, ‘train/Capture7’, ‘train/Capture8’, ‘train/Capture9’, ‘train/Capture10’, ‘train/Capture11’, ‘train/Capture12’, ‘train/Capture13’, ‘train/Capture14’, ‘train/Capture15’, ‘train/Capture16’, ‘train/Capture20’, ‘train/Capture22’, ‘train/Capture23’, ‘train/Capture24’, ‘train/Capture25’ for training. We exclude the pose sequences ‘0000_neutral_relaxed’, ‘0009_thumbtucknormal’,



Figure 7: Qualitative comparison with [41] on text-to-avatar application. The first column is the input reference images, and the rendered target views are displayed on the right side, respectively. The following two visual results are obtained by animating the reconstructed 3D hand avatar to different poses.



Figure 8: Qualitative comparison on HanCo [42]. The white box indicates the input view.

Method	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$
HandAvatar [3]	24.68	8.25	0.896
OHTA [41]	25.24	7.18	0.923
Ours	25.86	6.76	0.974

Table 3: Evaluation results on HanCo.

‘0019_alligator_closed’, ‘0029_indextip’, ‘0039_fingerspreadrigid’, ‘0048_index_point’, ‘0058_middelfinger’ to ensure a consistent experimental setting with [41]. For evaluation of one-shot hand avatars, we follow [3] to use sequence ‘test/Capture0/ROM03_RT_No_Occlusion’ with fixed skip steps. For each frame, we follow [3] and use the annotated detection box as the ground-truth hand region. The box is first adjusted to a square and enlarged by a factor of 1.3, after which the hand region is cropped and resized to a resolution of 256×256.

HanCo [42]. For quantitative experiments on HanCo dataset, we use sequence ‘0154’ with the cameras 0,3,5,6,7. We apply the provided MANO annotations, camera poses and hand masks of the dataset for one-shot reconstruction.

In-the-wild Data. For in-the-wild evaluation, we use whole-body images from COCO-Hand [20] and WHIM [23]. For each input image, we first use WiLoR [23] to detect the hand and determine its handedness. We then adopt the right hand for pose estimation, where [23] predicts the camera parameters and MANO annotations of the input images. We further use the corresponding pose estimation results to generate hand masks.

C.3 Quantitative Comparison.

To show the robustness of different subjects from the training set, we also perform the quantitative comparison on the HanCo dataset [42] with [3, 41], as they show robust one-shot reconstruction on InterHand2.6M. As shown in tab. 3, our method outperform previous works consistently in all metrics.

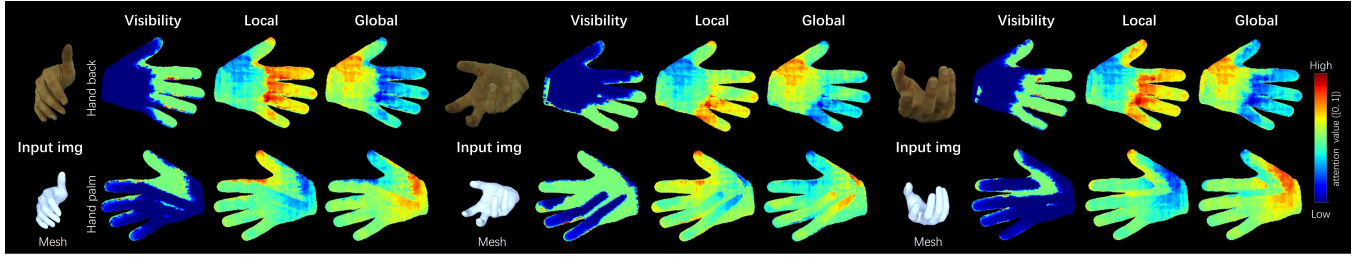


Figure 9: Points visibility and VPIA attention visualization.

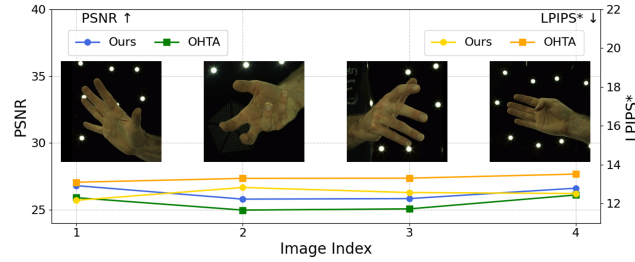


Figure 10: Robustness towards different input images. The images above the figure are the corresponding input images.

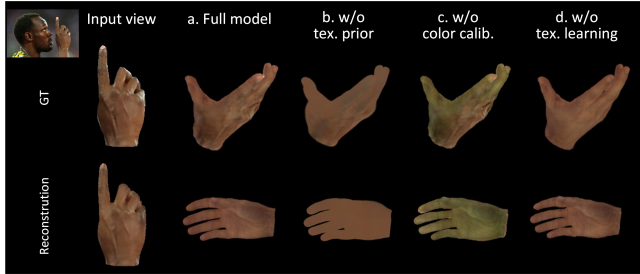


Figure 11: Visual results for ablation study with in-the-wild input.

#	Method	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$
a	Full Model	27.38	11.45	0.956
b	w/o Tex. prior	25.88	16.24	0.912
c	w/o Color Calib.	26.59	11.98	0.951
d	w/o Tex. Learning	27.02	12.18	0.953
e	w/o Regularization	27.08	12.28	0.954
f	w/o LoRA Finetune	27.11	12.06	0.954

Table 4: Ablation of one-shot reconstruction on InterHand2.6M. Tex. is short for texture, and calib. is short for calibration.

C.4 Qualitative Comparison

With respect to the qualitative comparison results on the HanCo dataset, fig. 8 highlights that our method maintains more consistent appearance across diverse hand poses and viewpoints. In contrast, HandAvatar [3] fails to recover plausible textures for unseen regions due to the incomplete information provided by single-view conditioning. OHTA [41] tends to produce noisy texture and unstable boundary transitions, especially around the occluded regions. By comparison, our method produces cleaner textures and smoother transitions under large viewpoint variations, leading to more coherent and visually plausible reconstructions.

Qualitative comparisons on in-the-wild images with OHTA [41] are shown in fig. 6. We further demonstrate the strong versatility in texture editing and text-to-avatar generation in fig. 1 and 7. For texture editing, we first edit the input image and obtain the corresponding edit mask. During one-shot reconstruction, we exclude the edited region when performing global color calibration, since the edited content may introduce colors that differ significantly from those of the hand and thus bias the global color statistics. We then run the texture learning stage for 1000 steps to adapt the reconstructed texture to the edited appearance and recover its high-frequency details. For text-to-avatar generation, we use Nano Banana with depth maps and text prompts as input for image generation. Subsequently, we perform one-shot reconstruction from these images with 1000 steps to better capture the complex details.

C.5 Quantitative Comparison for Robustness

Fig. 10 shows a per-image quantitative comparison between our method and OHTA [41] on different input images from InterHand2.6M. Our method achieves consistently better reconstruction quality, yielding higher PSNR and lower LPIPS across diverse examples. These stable improvements over varying hand poses and viewpoints demonstrate the robustness of the proposed approach.

C.6 Ablation Studies

Evidence routing of VPIA. We visualize the estimated visibility, a point-wise visibility probability obtained by sigmoid normalization to the z-buffer depth residuals, and the attention map in fig. 9. Visible regions in the input view generally show higher local attention, while unseen or self-occluded regions show higher global attention. This confirms that OASIS uses local image evidence for

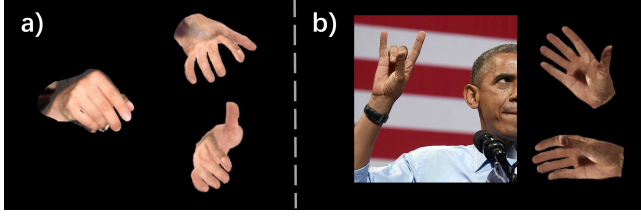


Figure 12: Failure cases.

visible regions and shifts to global context for occluded regions, which is consistent with the design of VPIA.

One-shot Reconstruction. To validate each component’s effectiveness, we perform ablation studies under the one-shot setting. The quantitative results are summarized in tab. 4 on the Inter-Hand2.6M dataset, and the qualitative results are performed on in-the-wild images, displayed in fig. 11. As shown in tab. 4 #b, removing texture priors and instead merely performing the one-shot strategy under the single-view setting leads to consistently worse reconstruction quality. We attribute this degradation to the fact that one-shot supervision is highly limited and under-constrained; without priors knowledge, the model lacks a strong initialization for unseen appearances (as shown in fig. 11 b), while full tuning substantially enlarges the optimization space and makes target fitting less stable. Without color calibration, the model suffers from the skin-color bias of the training data, which leads to a noticeable drop in PSNR as shown in tab. 11 #c. As seen in fig. 11 c, visible regions can still be learned from the input image, while unseen regions

are more strongly determined by the learned prior, retaining clear color discrepancies. This highlights the role of color calibration in reducing the training-domain color bias carried by the prior, which is essential for maintaining coherent appearance reconstruction across the whole hand. Omitting texture learning leads to a significant deterioration in LPIPS, as reported in tab. 4 #d, indicating a clear loss in perceptual fidelity. Qualitatively, fig. 11 d shows that the model can only rely on the prior network to preserve a coarse texture structure, while failing to reconstruct high-frequency details. This confirms that texture learning is crucial for injecting fine-grained image-specific cues beyond the generic prior, thereby enabling faithful one-shot appearance reconstruction.

D Limitations and Failure Cases

(a) *Dependency on hand estimators.* OASIS depends on off-the-shelf estimators [23] for pose, camera, handedness, and masks. Poor estimates, especially for challenging poses or extreme viewpoints, can lead to inferior texture modeling since one-shot personalization is optimized from an unreliable geometric initialization and observation. Although the learned hand prior provides robustness to moderate estimation errors, OASIS cannot fully handle highly inaccurate pose and camera estimates, as shown in fig. 12 (a).

(b) *Shadow/albedo ambiguity.* OASIS reconstructs observed appearance from single images and does not explicitly perform intrinsic decomposition; therefore, strong cast shadows may be baked into the texture (see fig. 12 (b)). This is a fundamental single-image inverse-rendering ambiguity, since material, geometry, and lighting are entangled. We will explore integrating material-aware decomposition into OASIS to separate illumination-dependent effects in future work.