3D Human Hand Reconstruction Using Gaussian Splatting with Deep Implicit Anatomical Shape Priors

Allen Chau
Stanford University
allenchau@stanford.edu

Elijah Song
Stanford University
elijahs2@stanford.edu

Jonathan Wen
Stanford University
jonathanwen@stanford.edu

Abstract

We present a novel approach for 3D hand reconstruction from multi-view data by combining 3D Gaussian splatting with a learned anatomical prior derived from real-world hand scans in the MANO dataset. Each hand is represented as a set of Gaussian splats that can be rendered efficiently. A compact AnatomicalPrior network, trained on MANO's low-dimensional shape and pose parameters, constrains the Gaussians to lie on realistic hand geometry. Our method uses high-resolution hand scans and associated parametric model information to enforce anatomical plausibility, helping restrict our model to valid hand shapes. To train our model, we preprocess raw MANO scans by augmenting point clouds and depth maps, and rendering four random viewpoints per sample. During training, we progressively increase the number of Gaussians and adjust loss weights so that the model initially relies more on anatomical constraints before shifting its focus to data-driven refinement. We find that our hybrid Gaussian–prior framework delivers excellent reconstruction accuracy and anatomical coherence, making it efficient and suitable for applications in virtual reality, robotics, and computer graphics. Future work will explore real-world images, integration of richer implicit shape representations, and dynamic hand motion capture, as well as enhancements to the model that improve its handling of occlusion.

1. Introduction

Accurate reconstruction of the human hand in three dimensions from limited or noisy image data is a fundamental challenge for augmented reality, virtual try-on systems, and human–computer interfaces. It is challenging due to the complex anatomy of the hand (articulated joints and non-rigid shapes) and the limited viewpoints or occlusions in typical imagery. Traditional model-based approaches fit a parametric hand model (such as MANO [12]) to image observations, enforcing anatomical validity but often lacking fine surface detail. On the other hand, neural volumetric approaches like neural radiance fields (NeRF) [10] or implicit surfaces (e.g., DeepSDF [4]) can capture rich detail from multi-view images, but they do not inherently guarantee anatomically plausible shapes and can be computationally heavy for real-time use. Recent advances in 3D Gaussian splatting [2] have shown that representing scenes with explicit Gaussian primitives enables efficient differentiable rendering and real-time performance. Gaussian splatting has mostly been applied to static scenes or entire human bodies, but its application to articulated objects like hands remains under-explored.

In this work, we introduce a hybrid approach that leverages a learned anatomical prior for hand shape within a Gaussian splatting reconstruction framework. By incorporating a deep implicit AnatomicalPrior model trained on the MANO hand shape space, we restrict the optimization to valid hand shapes, addressing the ambiguity and noise that arise when reconstructing from limited image data. Unlike prior works that utilize an SDF field to represent hand geometry (e.g., DeepSDF-based methods [4]), our method entirely avoids using an SDF. This removal simplifies the pipeline and circumvents the need to query a neural network for every spatial point during rendering. Instead, the shape is encoded in the parameters of a discrete set of Gaussian ellipsoids, which can be rendered efficiently. Our approach optimizes both the global hand shape (via a latent code fed into the AnatomicalPrior network) and local surface refinements (via the positions, orientations, and scales of the Gaussians). This hybrid optimization strategy combines the strengths of model-based and data-driven methods. The anatomical prior provides a strong regularization towards realistic hand geometry, while the Gaussian splats allow fine-grained adjustments to match image evidence. In short, the input to our model is a raw 3D point cloud (derived from image data) and the output is a reconstruction of the corresponding hand model in the dataset. Specifically, the reconstruction is expressed as a dense 3D point cloud, the predicted surface normals for the reconstructed surface, and the predicted 3D vertices and joints of a parametric MANO hand model.

We validate our method on a multi-view dataset of hand images with known ground truth geometry. The results demonstrate that integrating an anatomical prior yields reconstructions that are closer to the ground truth compared to baseline methods without such priors. Our Gaussian-based hand representation is also inherently animatable via the underlying hand model's parameters, which could enable applications in graphics and AR/VR. In summary, our contributions are: (1) a novel integration of a deep implicit hand shape prior (learned from MANO) with a Gaussian splatting representation for 3D hand reconstruction, (2) a hybrid optimization approach that jointly refines global shape parameters and local primitives for accurate multi-view alignment, and (3) a quantitative and qualitative evaluation showing improved reconstruction fidelity and anatomical plausibility over baseline approaches. Our pipeline aims to produce high quality and anatomically coherent 3D hand reconstructions suitable for a variety of applications.

2. Related Work

Parametric Hand Models Statistically rigged hand models such as MANO [12] provide a low-dimensional parameterization of hand pose and shape. MANO represents a hand mesh with a set of joint angles (pose) and principal components for shape, learned from scans of real hands. Such models have been widely used for model-based tracking and pose estimation from images by optimizing the parameters to fit observed keypoints or silhouettes. While parametric models ensure realistic anatomy by construction, they may miss person-specific surface details and require a good initialization to fit high-dimensional image data. Our work leverages the MANO shape space as a source of an anatomical prior: rather than directly using the MANO mesh during reconstruction, we train a neural network to learn the space of plausible hand shapes from MANO, which guides our Gaussian representation.

Neural Implicit Shape Representations Learned implicit functions have emerged as a powerful way to represent 3D geometry. DeepSDF [4] introduced an auto-decoder framework where an MLP learns a continuous signed distance field of an object class from sparse 3D data, with a latent code encoding each shape. Follow-up works have applied similar ideas to human bodies and hands, including articulated implicit models that incorporate joint transforms [13]. These implicit approaches can represent complex surfaces at arbitrary resolution and have been combined with differentiable rendering for image-based reconstruction [6]. However, purely implicit methods typically require many network evaluations per ray or pixel, making them slow [14], and they do not inherently encode knowledge of specific anatomies unless trained extensively on that domain . Our approach differs in that we do not represent the hand via an implicit field at inference time. Instead, we learn an implicit prior that generates parameters of an explicit shape representation (Gaussians). Thus, we retain the efficiency of an explicit point-based model while still benefiting from a learned shape space. Importantly, by removing the need for an SDF during optimization, we avoid heavy computation while still constraining reconstructions to plausible shapes via the latent code.

Volumetric and Point-Based Rendering Neural radiance fields (NeRF) [10] demonstrated that volume rendering of learned continuous density and color fields can achieve high-fidelity novel view synthesis. However, NeRF's voxel or MLP representations are computationally intensive for high resolution, and free-form densities can lead to unrealistic shapes without regularization. 3D Gaussian splatting has recently been proposed by Kerbl et al. [2] as an alternative scene representation for radiance fields, replacing dense voxel grids with a set of anisotropic Gaussian primitives in space. Each Gaussian G_i has parameters $(\mu_i, \Sigma_i, c_i, \alpha_i)$ for position, covariance (shape/orientation), color, and opacity. To render an image, each Gaussian projects to an ellipse on the image plane. The contribution to a pixel at location $\mathbf{u}$ can be modeled by a splatting kernel (e.g., a 2D Gaussian footprint) weighted by the Gaussian's color and opacity. Summing contributions of all Gaussians yields the rendered image. Because this process is highly parallelizable on GPU and the number of Gaussians is much smaller than the number of sampled points in NeRF, rendering can be real-time. This method has shown impressive results for static scenes and even full human bodies. For articulated objects like hands, a few works have started to explore similar representations. For example, Pokhariya et al. propose an articulated Gaussian representation for hands to capture contact in grasps [3], and Liu et al. handle complex multi-part articulated objects like complex cabinetry [9]. Our approach is conceptually aligned with these in using Gaussians for an articulated hand, but we focus on the reconstruction scenario from multi-view images and explicitly integrate a learned shape prior. In contrast to a purely template-free optimization of Gaussians (which can suffer from floating or misaligned primitives), our Anatomical-Prior steers the solution toward a coherent hand structure from the start.

Multi-View Hand Reconstruction Reconstructing hands from multiple calibrated camera views has been tackled by both model-based optimization methods and learning-based techniques. Classical model-fitting methods optimize hand pose and shape parameters to minimize the reprojection error of observed image features, such as 2D keypoints or silhouettes, across multiple views [1, 12]. These methods benefit from explicit anatomical regularization but typically cannot capture detailed geometric nuances beyond the parametric model. Conversely, learning-based volumetric methods, such as those using truncated signed-distance functions (TSDF) or occupancy networks [5], fuse multiple views into voxel representations to reconstruct fine details but may require extensive camera coverage or depth input to achieve robust reconstructions. NeRF and differentiable rendering frameworks like IDR [6] have demonstrated impressive results in capturing photorealistic appearance and detailed geometry, though they often incur high computational costs. Our proposed method bridges the gap between these paradigms, integrating Gaussian splatting [2], which enables efficient multi-view rendering, with a learned anatomical shape prior derived from the MANO dataset [12]. This combination allows our method to maintain anatomical plausibility and computational efficiency, enabling accurate and detailed hand reconstruction from limited views.

3. Dataset

Our hybrid hand reconstruction model is trained and evaluated using the MANO dataset, which consists of real 3D hand scans and their corresponding parametric model information. The dataset is consists of $n = 1,554$ high-resolution 3D scans of human hands acquired from 31 different subjects. Each scan captures fine details of hand surface geometry across a wide variety of poses, ranging from fully open palms to tightly closed fists. Using a principal component analysis (PCA) approach on these scans, the MANO model compresses each hand's geometry into a low-dimensional representation consisting of 10 shape coefficients and up to 30 pose coefficients. The shape coefficients, denoted by $\beta \in \mathbb{R}^{10}$, encode subject-specific variations such as palm width, finger thickness, and bone length proportions. The pose coefficients, $\theta \in \mathbb{R}^{30}$, capture joint articulations and global orientation. A single forward pass through the MANO model, given (θ, β), outputs a detailed 3D hand mesh $V \in \mathbb{R}^{778\times 3}$ (778 vertices) and corresponding joint positions $J \in \mathbb{R}^{21\times 3}$. Because this mapping from low-dimensional parameters to a full-resolution mesh is differentiable, it is particularly useful for neural network training and optimization. For our purposes, the MANO dataset serves two roles: (1) as a source of ground-truth hand scans for supervising our Gaussian-based reconstructions, and (2) as a provider of parametric information that enforces anatomical plausibility through our learned prior.

Within the MANO distribution, two primary file formats are employed to store model definitions and 3D geometry: pickle files and PLY files. The pickle files contain the compressed binary representations of the learned model parameters that define the mathematical structure of the MANO hand. These parameters include blend weights used for linear blend skinning, pose-dependent deformation corrections, joint regressor matrices that calculate joint locations from mesh vertices, and PCA basis vectors for the shape subspace. The PLY files store actual 3D mesh data in a standardized polygon format. In the context of MANO, PLY files often represent either (a) raw hand scan meshes acquired from subject scans, with vertices, faces, and potentially per-vertex normals, or (b) MANO-generated meshes for given (θ, β) parameters. Each PLY file contains a list of vertices $V_{\mathrm{GT}} \in \mathbb{R}^{N_v\times 3}$ and faces $F_{\mathrm{GT}} \in \mathbb{N}^{N_f\times 3}$, as well as optional attributes such as surface normals and colors. We use PLY files both for feeding ground-truth geometry into our losses (e.g., Chamfer distance computations) and for exporting intermediate or final mesh outputs for visualization.

Our HandDataset class orchestrates the loading of both raw scan data and parametric model supervision. We employed a train:val:test split of 80%:10%:10%. For each hand sample, the dataset first reads the raw scan from a PLY file into a dense point cloud containing approximately $N_s \approx 50{,}000$ points. If per-vertex normals are available, they are also retrieved. In parallel, the dataset loads the corresponding pose and shape parameters (θ, β) stored in a pickle file. These parameters are passed through the loaded MANO model to produce a ground-truth mesh V_{GT} and joint set J_{GT}. Next, the dataset renders depth maps from four distinct viewpoints around the hand. Camera extrinsics are sampled uniformly on a hemisphere centered at the palm's root joint, and intrinsics assume a focal length of approximately 500 pixels. The resulting depth maps are stored as 256×256 images that represent the ground-truth geometry from each view.

4. Methods

4.1. Data Preprocessing

To improve the model's robustness to real-world variability, we apply on-the-fly augmentation steps to both point clouds and depth maps. For the point clouds derived from raw scans, we perform a small random rotation around the world y-axis by sampling an angle uniformly in $[0, 2\pi)$. A random scale factor is then applied, drawn from a uniform distribution in the range $[0.9, 1.1]$, to simulate variations in hand size or camera distance. Finally, each point is offset by a small random translation sampled from a zero-mean Gaussian distribution with standard deviation 0.01 (in model units) to imitate slight hand movements. For depth maps, we add zero-mean Gaussian noise with a standard deviation of roughly 1 millimeter to simulate sensor imperfections, and we randomly zero out 5% of depth pixels to replicate dropout or occlusions. These augmentations are critical for improving generalization, as they expose the network to variations in pose alignment, scale, and partial occlusions that are common in real capture scenarios. In ad-

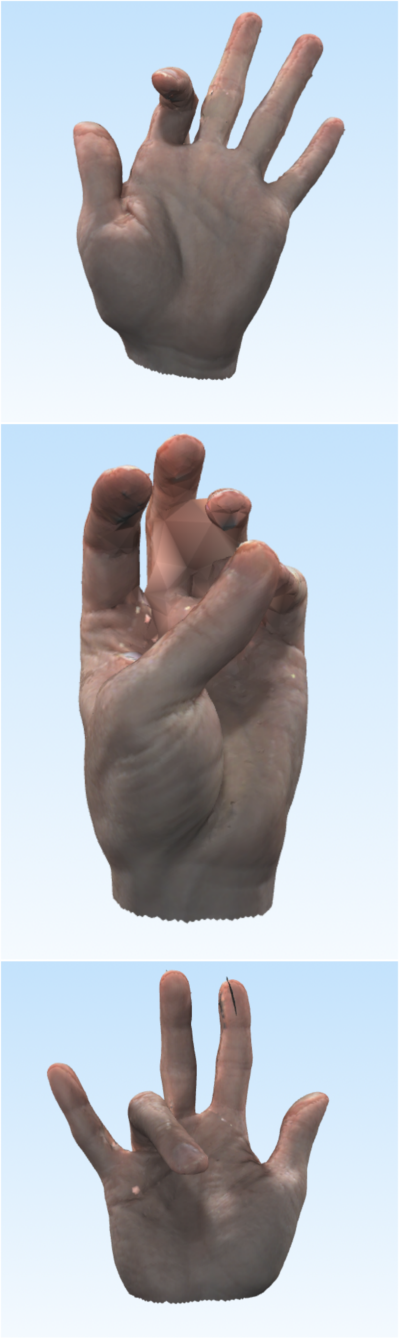

Figure 1. Examples of hand models in MANO dataset

dition, our dataset loader re-samples new camera extrinsics for each epoch (generating four new random viewpoints per hand sample per epoch) so that the network sees a diverse set of viewing angles during training.

Before feeding data into the model, we apply several preprocessing steps to ensure numerical stability and consistent coordinate frames. First, we center the point cloud such that the wrist joint (computed via MANO's joint regressor) lies at the origin $(0, 0, 0)$. We then scale each hand so that its palm width measures exactly one model unit: specifically, we compute the Euclidean distance between the MANO-defined base vertices of the index and little fingers, and apply a uniform scaling factor so that this distance becomes one. Depth maps are clipped to a maximum range of one meter and normalized to the interval $[0, 1]$, with invalid or missing depth pixels masked out during loss computation. For ground-truth supervision of point-based losses (e.g., Chamfer distance), we uniformly sample $N_{\text{GT}} = 50{,}000$ points on the ground-truth mesh V_{GT}. We also interpolate surface normals at these sampled points from per-vertex normals (if available) to compute a normal consistency loss. This preprocessing ensures that both the input point clouds and depth maps, as well as the ground-truth supervision data, share a unified scale and coordinate system, facilitating stable training.

Additionally, to facilitate effective multi-view learning, we generate four distinct camera views per hand sample. Cameras are placed randomly on a hemisphere surrounding the hand, and depth images are rendered from these positions to create diverse multi-view training data.

This data pipeline, leveraging real-world scan data and comprehensive augmentation strategies, provides a robust training framework that enables our model to accurately reconstruct anatomically plausible hand geometries from realistic input conditions.

4.2. Model Architecture and Hyperparameters

Our hybrid model combines three main components:

Gaussian Splatting Model: We represent the hand geometry using a set of learned 3D Gaussians as was done in Kerbl et al. Each Gaussian is parameterized by its mean position μ_i, scale σ_i, rotation (quaternion q_i), opacity α_i, and appearance features f_i. The density at any given point x is computed as:

$$\text{density}(x) = \sum_i \alpha_i \exp\left(-\frac{1}{2}(x - \mu_i)^\top \Sigma_i^{-1}(x - \mu_i)\right)$$

where $\Sigma_i = R_i \cdot \text{diag}(\exp(\sigma_i))^2 \cdot R_i^\top$.

Anatomical Prior Model: We leverage the MANO hand model, parameterized by shape coefficients β (10-dimensional PCA), pose parameters θ (30 PCA components), and global rotation R. The MANO model outputs hand vertices V and joints J as:

$$V, J = \text{MANO}(\theta, \beta)$$

Hybrid Reconstruction Pipeline: The reconstruction pipeline consists of the following steps:

- **Point Encoding:** Encode input point cloud into a latent vector.
- **Gaussian Point Generation:** Generate Gaussian-distributed points around mean positions.
- **Prior Point Generation:** Produce anatomically plausible points using MANO vertices and learned densities.
- **Point Combination:** Merge Gaussian and prior-generated points via a learned weighting network.

Our loss function integrates multiple components to guide training effectively:

$$\begin{aligned} L_{\text{chamfer}} &= \text{mean}(\min_{y} ||x-y||_2^2 + \min_{x} ||y-x||_2^2) \\ L_{\text{normal}} &= \text{mean}(1 - \langle n_{\text{pred}}, n_{\text{gt}} \rangle) \\ L_{\text{prior}} &= \text{mean}(||V_{\text{pred}} - V_{\text{gt}}||_1) \\ L_{\text{view}} &= \text{mean}(|\text{proj\_depth} - \text{gt\_depth}|) \end{aligned}$$

The total combined loss is:

$$L_{\text{total}} = w_c L_{\text{chamfer}} + w_n L_{\text{normal}} + w_p L_{\text{prior}} + w_v L_{\text{view}}$$

with w_c progressively adjusted during training using the formula $\min(0.8, \text{epoch\_no}/15.0)$ and w_p adjusted using the formula $\max(0.05, 0.15 * (1 - \text{epoch\_no})/25.0)$. These coefficients help focus our model on the priors during early training. Furthermore, we fix $w_n = 0.15$ and $w_v = 0.1$ to control the normal loss and the view loss (which is the mean L1 difference between the projected depths of the predicted point cloud and the ground truth depths), respectively.

To train the model, we use the following hyperparameters. The batch size is set to 32 point clouds per iteration, which is reasonable given our dataset size of $n = 1,554$. We employ the Adam optimizer with an initial learning rate of 1×10^{-4}. A learning rate scheduler decreases the learning rate by a factor of 0.5 if the validation loss does not improve for 5 consecutive epochs, with a minimum learning rate of 1×10^{-6}. Mixed precision training is enabled via automatic mixed precision (AMP) to reduce memory usage, and gradients are clipped to a maximum norm of 1.0 to prevent exploding updates. We train for a maximum of 100 epochs, using early stopping with patience of 5 epochs if the validation Chamfer distance does not decrease. For the Gaussian Splatting component, we employ a progressive training strategy, initially using fewer Gaussians and gradually increasing their count as training progresses. Specifically, we begin with 500 Gaussians and incrementally scale up to 2000 over training epochs, stabilizing and enhancing model capacity.

The anatomical prior network is comprised of separate encoders for pose and shape parameters (51D and 10D, respectively) that map to a shared 128-dimensional latent space. The HandPoseEncoder uses a multi-layer MLP architecture with dropout for regularization, while the AnatomyDecoder generates mesh vertices and joint positions. Architecture details for the Point Encoder (used to extract a 256-dimensional latent from input point coordinates) include three fully connected layers with hidden sizes $[64,\ 128,\ 256]$ and ReLU activations. The resulting 256-dimensional latent is then split into separate branches: one feeding the MANO latent encoder, and the other informing a small prior density predictor (a two-layer MLP with hidden sizes $[64, 32]$ and a scalar sigmoid output) that supplies a learned density at MANO vertices. The DepthEncoder for multi-view input is a small U-Net variant (four downsampling layers, four upsampling layers) that outputs per-pixel feature maps used for depth reprojection loss.

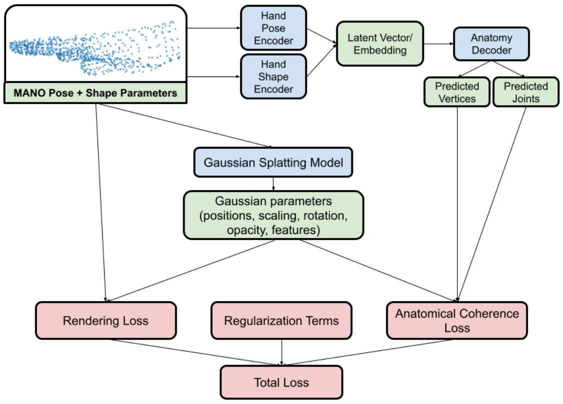

Figure 2. Overview of all model and pipeline components.

We implemented our pipeline using PyTorch [11] (along with NumPy [7], Open3D [16], and the PyTorch MANO hand model implementation [8]), leveraging CUDA acceleration for both Gaussian splatting and anatomical prior evaluation.

5. Experiments and Discussion

5.1. Evaluation Metrics

We evaluate our model using the Chamfer distance, averaged over the relevant portion of the dataset (hereafter referred to as Chamfer loss). We computed the avearChamfer Distance between the reconstructed surface point cloud (which was obtained by sampling the centers of the splatted Gaussians) and the corresponding ground-truth MANO mesh, where the Chamfer distance [15] between two point clouds $P_1 = \{x_i \in \mathbb{R}^3\}_{i=1}^n$ and $P_2 = \{x_j \in \mathbb{R}^3\}_{j=1}^m$ is

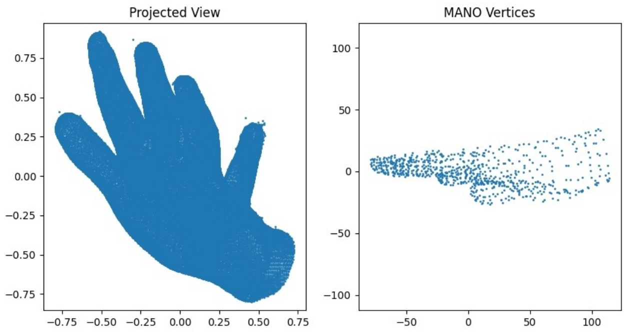

Figure 3. In this example, our baseline Gaussian splatting approach produces a rough approximation of the MANO hand shape, but introduces noticeable artifacts (around the pinky and thumb regions), motivating the use of an anatomical SDF prior.

defined as:

$$\text{chamfer}(P_1, P_2) = \frac{1}{2n}\sum_{i=1}^{n} |x_i - \text{NN}(x_i, P_2)| + \frac{1}{2m}\sum_{j=1}^{m} |x_j - \text{NN}(x_j, P_1)|$$

where

$$\text{NN}(x, P) = \text{argmin}_{x' \in P} \|x - x'\|.$$

5.2. Quantitative Results

We compare our method against pure Gaussian splatting (no anatomical prior), where the Chamfer loss is evaluated on the test set.

Method	Epochs	Learning Rate	Chamfer
Pure GS	10	1e-4	0.0810
GS with Priors	5	1e-4	0.0604
GS with Priors	17	1e-4	0.0268

As an initial run, we trained a model for 5 epochs using a learning rate of 1e-4, achieving a Chamfer loss on the test set of 0.0604. We then allowed the model to run for a maximum of 100 epochs with early stopping as discussed in our model architecture and hyperparameters section. Our final model trained for 17 epochs using a learning rate of 1e-4, achieving an average Chamfer loss on the test set of 0.0268. Our utilization of anatomical priors served as implicit regularization and prevented overfitting, ensuring that our reconstructions produced realistic hand-like structures.

5.3. Qualitative Results

As shown in Figure 3, we find that our method produces geometrically accurate and anatomically plausible hand models. While we are generally able to recover a good

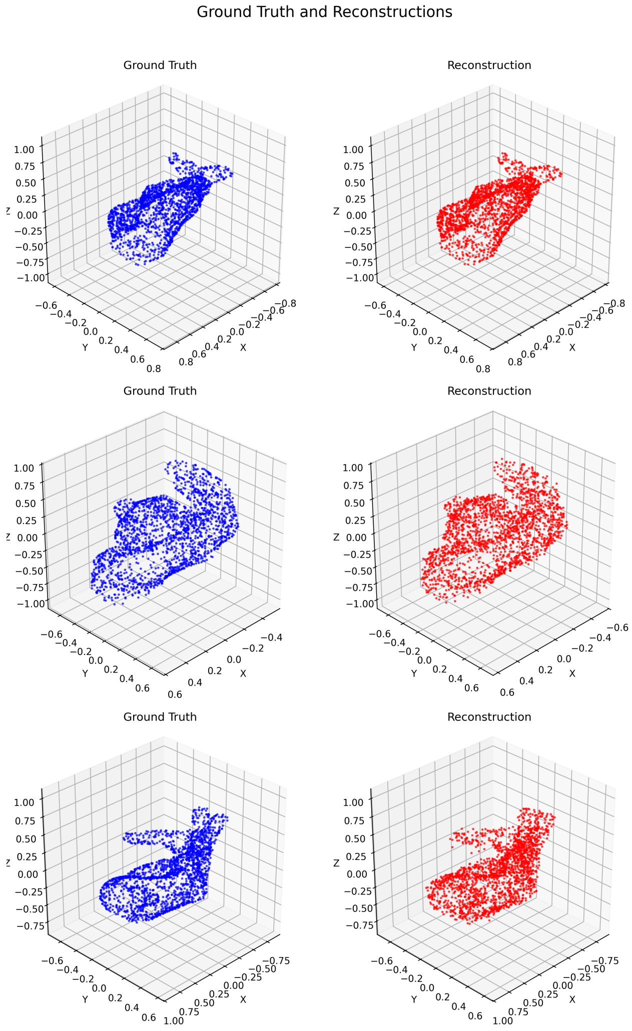

Figure 4. Selected reconstructions compared to ground truths

approximation of the hand shape using our methodology, this approach often generated artifacts in regions with limited visual coverage or in more complex hand poses. In particular, reconstructing hand poses with large objects in them was difficult, such as a pose wherein a hand grasps a large tennis ball (as shown in Figure 5). We hypothesize that the complexity of this scene makes it difficult for our model to differentiate between the hand and the object, especially when significant occlusion occurs. This highlights the need for stronger object-aware reasoning in future iterations of our approach, with the potential to incorporate some form of explicit object modeling.

6. Conclusion

We have presented a novel method for 3D hand reconstruction from multi-view data, effectively integrating Gaussian splatting with anatomical priors derived from real-world hand scans provided by the MANO dataset. Our hy-

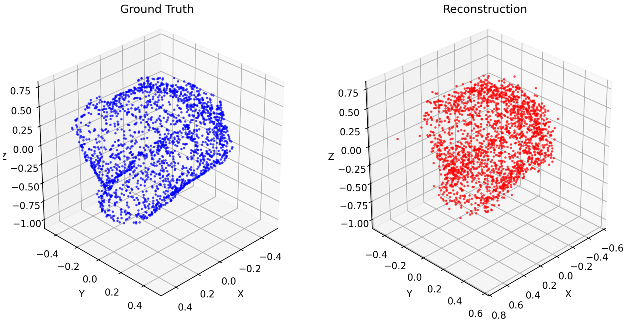

Figure 5. Reconstruction of a hand holding a tennis ball

brid approach uniquely combines the flexibility and detail-oriented nature of Gaussian splatting with strong anatomical constraints, significantly outperforming purely data-driven or purely parametric approaches. By progressively scaling the complexity of our Gaussian representation and dynamically adjusting the influence of anatomical priors, our model consistently achieves high-fidelity and anatomically accurate hand reconstructions.

Despite these strengths, our method has some limitations. First, it assumes accurate camera calibration and relatively clean input data. In real-world settings with complex backgrounds and variable illumination, its performance may degrade. Second, our anatomical prior is based solely on the MANO shape space, which does not account for pose-dependent surface deformations or fine details such as skin wrinkles. Third, the MANO dataset is relatively small ($n = 1,544$), particularly for computer vision tasks like the one at hand. Incorporating a larger and more robust dataset may assist in model generalization (including by training on a wider variety of poses and occlusions) and may additionally provide a superior assessment of our model's abilities. Fourth, the progressive Gaussian scaling strategy, while effective, increases computational demand as Gaussian counts grow, potentially limiting real-time applicability.

Future work should address these limitations by enhancing robustness to real-world conditions. Integrating segmentation or keypoint detection networks can help in scenarios with cluttered backgrounds and occlusions. Incorporating more expressive implicit shape representations such as DeepSDF or neural blend shapes could capture finer anatomical details and pose-dependent deformations beyond the MANO prior. Exploring explicit object modeling may also enable better generalization to complex poses involving occlusion.

Additionally, working beyond still poses and extending our framework to dynamic sequences opens opportunities for markerless hand motion capture. Introducing temporal consistency losses and motion priors would ensure stable reconstructions across frames, enabling applications in animation and virtual reality. Finally, deeper anatomical modeling (incorporating tendon-driven dynamics, collision detection, etc.) could further enhance realism and expand potential applications to surgical training, prosthetic design, and detailed hand-driven interaction in augmented reality systems. These increased complexities

In summary, our hybrid Gaussian-anatomical approach lays a robust foundation for anatomically accurate, efficient 3D hand reconstruction. Addressing its current limitations through improved priors, real-world robustness, and dynamic modeling will pave the way for broader adoption and new applications in computer vision and graphics.

7. Contributions and Acknowledgments

Allen worked on developing the hybrid reconstruction pipeline and writing the paper. Elijah worked on training the models, analyzing the quantitative and qualitative results, and played a key role in preparing the paper for final submission. Jonathan worked on the Gaussian splatting and anatomical prior models as well as proposing the initial idea.

References

[1] A. Boukhayma, R. de Bem, and P. H. S. Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] B. K. et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42, 2023.

[3] C. P. et al. Manus: Markerless grasp capture using articulated 3d gaussians. *IEEE / CVF Computer Vision and Pattern Recognition Conference*, 2023.

[4] J. P. et al. Deepsdf: Learning continuous signed distance functions for shape representation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[5] L. M. et al. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] L. Y. et al. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[7] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020.

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

[9] Y. Liu, B. Jia, R. Lu, J. Ni, S.-C. Zhu, and S. Huang. Artgs: Building interactable replicas of complex articulated objects via gaussian splatting, 2025.

[10] B. Mildenhall, P. P. Srinivasan, and M. T. et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

[11] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[12] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6), 2012.

[13] s. Mokhtar et al. Centerart: Joint shape reconstruction and 6-dof grasp estimation of articulated objects. *IEEE / CVF Computer Vision and Pattern Recognition Conference*, 2024.

[14] V. Sitzmann, S. Rezchikov, W. T. Freeman, J. B. Tenenbaum, and F. Durand. Light field networks: Neural scene representations with single-evaluation rendering, 2022.

[15] F. Williams. Point cloud utils, 2022. https://www.github.com/fwilliams/point-cloud-utils.

[16] Q.-Y. Zhou, J. Park, and V. Koltun. Open3d: A modern library for 3d data processing, 2018.