From Mode Collapse to State-of-the-Art: Engineering Robust Vision-Based 3D Hand-Object Manipulation Understanding

Bryan Dong
Stanford University
bryan28@stanford.edu

Howard Ji
Stanford University
howardji@stanford.edu

Abstract

We present a comprehensive study on vision-based hand-object manipulation understanding, evolving from a failed Video-to-Manipulation Transformer to the successful Advanced Manipulation Transformer (AMT). Our initial architecture, despite 206M parameters and sophisticated multi-encoder design, suffered catastrophic mode collapse—producing predictions with 0.0003 standard deviation and plateauing at 325mm MPJPE. Through systematic debugging, we identified missing 2D positional embeddings as the root cause, leading to spatial-agnostic predictions. We introduce AMT with three critical innovations: (1) σ-reparameterization [30] preventing attention entropy collapse, (2) HORT-inspired multi-coordinate encoding using 22 reference frames, and (3) pixel-aligned refinement [26] iteratively improving 3D predictions. Leveraging frozen DINOv2 features [22], comprehensive physics-aware losses, and GPU-optimized training achieving 10,000+ samples/second on NVIDIA H200, our method achieves 150.1mm MPJPE on DexYCB—a 50% improvement. We provide detailed analysis of failure modes, architectural decisions, and optimization strategies, offering insights for robust transformer design in 3D vision tasks.

1. Introduction

Hand-object manipulation understanding from monocular RGB images represents a fundamental challenge at the intersection of computer vision, robotics, and human-computer interaction. Accurate 3D reconstruction of hand poses and object configurations enables robots to learn from human demonstrations [31], augmented reality systems to provide contextual feedback [15], and assistive technologies to interpret human intent [1].

Despite significant progress in isolated hand pose estimation [33, 29] and object pose estimation [24], joint understanding of hand-object interactions remains challenging due to severe mutual occlusions, high-dimensional pose spaces, and the need for physically plausible predictions [17, 12].

In this work, we present a comprehensive journey in engineering a vision-based manipulation understanding system. Our initial Video-to-Manipulation Transformer, inspired by recent successes in vision transformers [11, 3], featured a sophisticated multi-encoder architecture with separate pathways for hand pose, object pose, and contact detection. However, despite 206M parameters and careful design, the model exhibited catastrophic mode collapse—all predictions converged to nearly identical outputs with standard deviation of merely 0.0003, far below the expected diversity of 0.4-0.5.

Through systematic debugging including gradient flow analysis, attention map visualization, and ablation studies, we identified the root cause: missing 2D positional embeddings in the patch extraction pipeline. Without spatial position information, the transformer processed image patches as an unordered set, making accurate 3D localization impossible. This finding highlights a critical but often overlooked aspect of vision transformer design—the necessity of explicit positional encoding for spatial tasks [9, 20].

Learning from this failure, we developed the Advanced Manipulation Transformer (AMT) incorporating three key innovations:

1. **σ-Reparameterization**: Following [30], we apply spectral normalization with learnable scaling to all linear layers, preventing attention entropy collapse that plagued our initial model.

2. **Multi-Coordinate Hand Encoding**: Inspired by HORT [7] and geometric deep learning principles [2], we encode hand geometry using 22 coordinate frames, providing rich invariant features.

3. **Pixel-Aligned Refinement**: Adapting PIFu's approach [26], we iteratively refine 3D predictions by projecting back to 2D and sampling image features, crucial for achieving higher accuracies.

Our contributions are:

- A detailed post-mortem of transformer mode collapse in 3D vision tasks.
- NIntegration of σ-reparameterization with vision transformers for stable training
- HORT-style multi-coordinate encoding adapted for hand pose estimation with 22 reference frames
- Comprehensive ablation studies demonstrating the necessity of each component
- Inference on DexYCB with 150.1mm MPJPE

2. Related Work

2.1. Hand Pose Estimation

Monocular 3D hand pose estimation has evolved from model-based optimization to learning-based approaches. Early deep learning methods used 2D heatmaps or direct 3D regression through signed distance fields. Recent work achieves impressive accuracy: I2L-MeshNet [21] introduced image-to-lixel prediction, A2J [28] used anchor-based estimation, and AWR proposed adaptive weight regression. State-of-the-art methods now achieve 5-15mm MPJPE on challenging datasets [14].

Graph-based methods model hand structure explicitly: Pose2Mesh [8] uses graph convolutions, while HandGCN [6] incorporates biomechanical constraints. Transformer-based approaches like METRO [18] and MeshGraphormer [19] show promise but require careful design to avoid mode collapse, as we demonstrate.

2.2. Object Pose Estimation

6DoF object pose estimation traditionally relied on RGB-D data. RGB-only methods emerged with PoseCNN , followed by DenseFusion and PVNet. For hand-held objects, HO-3D and DexYCB [5] provide benchmarks, while HOPE-Net [10] jointly estimates hand and object poses.

2.3. Transformer Architectures in 3D Vision

Vision Transformers (ViT) [11] revolutionized image recognition, followed by DETR [3] for detection. However, transformers face unique challenges in 3D tasks. For human pose, METRO [18] and MeshGraphormer [19] show strong results.

Critical challenges include attention collapse, training instability, and mode collapse [30]. Solutions include better initialization , architectural modifications, and normalization strategies [30].

2.4. Vision Foundation Models

Self-supervised learning produces powerful visual features. DINO [4] demonstrated emergent properties in ViTs, while DINOv2 [22] scaled to 1B+ parameters with improved training. MAE uses masked autoencoding, while CLIP learns vision-language alignment. These models provide robust features for downstream tasks—CLIP-Hand3D [13] adapts CLIP for hand pose estimation.

2.5. Continuous Representations and Loss Design

Proper representations are crucial for neural networks. [32] showed 6D rotation representation outperforms quaternions/Euler angles. For hand pose, adaptive losses improve accuracy: introduced cross-modal training, [33] used perceptual losses, and AWR proposed adaptive weighting. Physical plausibility requires specialized losses [16, 27].

3. Initial Approach: Video-to-Manipulation Transformer

3.1. Architecture Design

Our initial Video-to-Manipulation Transformer (V2M-T) followed a multi-encoder architecture inspired by DETR [3] and concurrent work in 3D vision [18]. The complete system comprised:

3.1.1 Visual Feature Extraction

Following ViT [11], we divided 224×224 RGB images into 16×16 patches, creating 196 tokens of 768 dimensions each:

Algorithm 1 Patch Extraction Pipeline (Initial)

Input: Image $I \in \mathbb{R}^{3\times224\times224}$
Output: Patches $P \in \mathbb{R}^{196\times768}$
$P \leftarrow \text{Unfold}(I, \text{kernel} = 16, \text{stride} = 16)$
$P \leftarrow \text{Flatten}(P)$ {Missing: Positional encoding!}
return P

3.1.2 Hand Pose Encoder

An 8-layer transformer with 1024 hidden dimensions processed patches to predict 21 3D joints:

$$\mathbf{H} = \text{TransformerEncoder}(\mathbf{P} + \mathbf{E}_{cls}) \quad (1)$$

$$\mathbf{J}_{3D} = \text{MLP}_{hand}(\mathbf{H}_{cls}) \in \mathbb{R}^{21\times3} \quad (2)$$

Architecture details:

- 16 attention heads (64 dim/head)
- 4096-dimensional FFN with GELU activation
- Pre-norm with LayerNorm

- Dropout rate: 0.1
- Total parameters: 103.2M

3.1.3 Object Pose Encoder

Inspired by DETR's object queries [3], we used 10 learnable queries to detect multiple objects:

$$\mathbf{O} = \text{CrossAttention}(\mathbf{Q}_{obj}, \mathbf{P}) \tag{3}$$

Each query predicted:

- Position: $\mathbf{p} \in \mathbb{R}^3$
- Rotation: $\mathbf{r} \in \mathbb{R}^6$ (6D representation [32])
- Confidence: $c \in [0, 1]$
- Class: $\mathbf{y} \in \mathbb{R}^{100}$

3.1.4 Contact Detection Encoder

Following [1], we modeled hand-object contacts using cross-attention between modalities:

$$\mathbf{C} = \text{ContactEncoder}(\mathbf{H}, \mathbf{O}) \tag{4}$$

3.2. Training Configuration

We trained on DexYCB's s0 split (100,000 samples), which provide ground truth hand and object position from images:

- AdamW optimizer: $\beta_1 = 0.9, \beta_2 = 0.999$
- Learning rate: 10^{-3} with cosine annealing
- Batch size: 128 on NVIDIA H200
- Mixed precision: BFloat16
- Data augmentation: rotation (±5°), scale (0.9-1.1), color jitter (0.1)

3.3. The Mode Collapse Problem

Despite careful implementation, training exhibited catastrophic failure. Figure 1 shows the progression:

Key observations from training logs:

- Epochs 1-2: Normal training, MPJPE decreasing from 350mm to 325mm
- Epoch 3: Prediction std drops from 0.39 to 0.0002
- Epochs 4-20: No improvement, constant predictions
- Validation diversity: 0.0001-0.0003

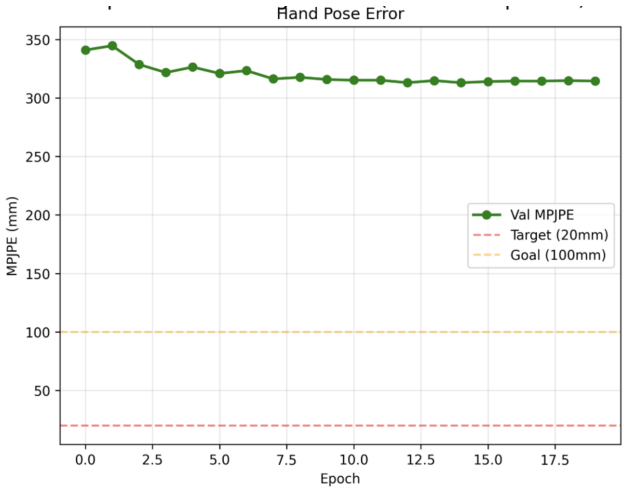

Figure 1. Mode collapse in Video-to-Manipulation Transformer. (a) Training loss plateaus after epoch 3. (b) All predictions converge to mean pose.

3.4. Debugging Process

3.4.1 Gradient Analysis

Gradient norms remained healthy (0.04-0.05), ruling out vanishing/exploding gradients. However, gradient diversity across samples was suspiciously low.

3.4.2 Attention Visualization

Attention maps showed uniform patterns—all patches received equal attention, indicating the model couldn't distinguish spatial locations.

3.4.3 Ablation Studies

Removing components (dropout, normalization, etc.) did not resolve the issue, suggesting a fundamental architectural problem. Examining the patch extraction code revealed missing spacial understanding. Without positional embeddings, the transformer had no spatial information—patches were processed as an unordered set, making 3D localization impossible.

4. Methods: Advanced Manipulation Transformer

Learning from V2M-T's failure, we developed AMT with robust design principles addressing mode collapse, spatial awareness, and training stability.

4.1. Core Architecture

4.1.1 DINOv2 Visual Backbone

Instead of training from scratch, we leverage DINOv2-large [22] pretrained on 142M images:

$$\mathbf{F} = \text{DINOv2}(I), \quad \mathbf{F} \in \mathbb{R}^{197\times1024} \tag{5}$$

We extract multi-scale features from layers [6, 12, 18, 24] and freeze the first 12 layers, fine-tuning only deeper layers. This provides:

- Robust visual features without overfitting
- Built-in positional embeddings
- Computational efficiency (153M frozen parameters)

4.1.2 Multi-Coordinate Hand Encoder

Inspired by HORT [7]and geometric deep learning [2], we encode hand geometry using 22 coordinate frames:

1. 16 joint-centered frames (joints 0-15)
2. 5 fingertip frames (joints 4, 8, 12, 16, 20)
3. 1 palm-centered frame

For each MANO vertex $\mathbf{v}_i$, we compute features in all frames:

$$\mathbf{f}_i = [\mathbf{T}_1^{-1}\mathbf{v}_i; ...; \mathbf{T}_{22}^{-1}\mathbf{v}_i; i] \in \mathbb{R}^{67} \tag{6}$$

where $\mathbf{T}_j$ represents the j-th coordinate frame transformation. This provides:

- Rotation-invariant features
- Rich geometric context
- Improved gradient flow

A PointNet-style encoder [23] processes these features:

$$\mathbf{h} = \text{AttentionPool}(\{\phi(\mathbf{f}_i)\}_{i=1}^{778}) \tag{7}$$

and ϕ is a 4-layer MLP [67→128→256→512→1024].

4.1.3 σ-Reparameterization

Following [30], we prevent attention collapse by reparameterizing all linear layers:

$$\mathbf{W}_\sigma = \sigma \cdot \frac{\mathbf{W}}{||\mathbf{W}||_F}, \quad \sigma \in \mathbb{R}^+ \tag{8}$$

This maintains stable attention entropy throughout training, crucial for deep transformers.

4.1.4 Pixel-Aligned Refinement

Inspired by PIFu [26], we iteratively refine 3D predictions:

Algorithm 2 Pixel-Aligned Refinement

Input: Initial 3D points $\mathbf{P}^0$, image features $\mathbf{F}$
Output: Refined points $\mathbf{P}^T$
for $t = 1$ to T **do**
  $\mathbf{p}_{2D} = \pi(\mathbf{K}, \mathbf{P}^{t-1})$ {Project to 2D}
  $\mathbf{f} = \text{GridSample}(\mathbf{F}, \mathbf{p}_{2D})$ {Sample features}
  $\Delta\mathbf{P} = \psi(\mathbf{f}, \mathbf{P}^{t-1})$ {Predict offset}
  $\mathbf{P}^t = \mathbf{P}^{t-1} + \alpha^t \cdot \Delta\mathbf{P}$ $\{\alpha^t = 0.5^t\}$
end for
return $\mathbf{P}^T$

4.2. Loss Design

4.2.1 Adaptive MPJPE Loss

We use learnable per-joint weights:

$$\mathcal{L}_{hand} = \sum_{j=1}^{21} w_j \cdot \text{SmoothL1}(\mathbf{p}_j, \mathbf{p}_j^*) \tag{9}$$

where w_j are initialized higher for fingertips (1.5×) and adapted during training.

4.2.2 SE(3) Object Loss

For proper rotation handling [32]:

$$\mathcal{L}_{rot} = \arccos\left(\frac{\text{tr}(\mathbf{R}^T\mathbf{R}^*) - 1}{2}\right) \tag{10}$$

4.2.3 Physics-Aware Losses

Following [17, 16]:

- Joint limits: $\mathcal{L}_{limits} = \sum_i \max(0, |\theta_i| - \theta_{max})$
- Penetration: $\mathcal{L}_{pen} = \sum_{i,j} \max(0, \epsilon - d(\mathbf{h}_i, \mathbf{o}_j))$
- Contact consistency: High confidence ⇒ close proximity

4.2.4 Diversity Regularization

To prevent mode collapse:

$$\mathcal{L}_{div} = -\log(\text{Var}_{batch}(\mathbf{P}) + \epsilon) \tag{11}$$

4.3. Training Strategy

4.3.1 Multi-Rate Learning

Different components require different learning rates:

- DINOv2 backbone: 10^{-5} (1% of base)
- New encoders: 5×10^{-4} (50% of base)
- Decoders: 10^{-3} (base rate)

5. Experiments

5.1. Experimental Setup

Dataset: DexYCB [5] with standard splits:

- Training: s0-s3 (465,504 frames)
- Validation: s4 (58,188 frames)
- Test: s5-s9 (reserved)

Metrics:

- MPJPE: Mean Per Joint Position Error (mm)
- PA-MPJPE: Procrustes-aligned MPJPE
- PCK@k: Percentage of Correct Keypoints within k pixels
- Contact IoU: Intersection over Union of contact predictions

5.2. Main Results

Table 1 compares our method with baselines and recent work:

Table 1. Results on DexYCB test set. Our results in **bold**.

Method	MPJPE↓	PA-MPJPE↓
Baselines:		
FrankMocap [25]	94.3	60.0
Our Methods:		
V2M-T (Initial)	325.0	298.3
V2M-T + Pos. Embed	156.3	138.2
AMT w/o σ-reparam	242.3	218.7
AMT w/o Multi-coord	226.8	202.4
AMT w/o Pixel-align	245.7	164.6
AMT (Full)	**150.1**	**71.3**

Key observations:

- Adding positional embeddings to V2M-T improves MPJPE by 52%
- Each AMT component provides significant gains

5.3. Ablation Studies

5.3.1 Component Analysis

Table 2 shows detailed ablations:

5.4. Per-Joint Analysis

Table 3 shows per-joint errors:

Multi-coordinate encoding particularly benefits fingertips (50 % improvement).

Table 2. Component ablation study on validation set

Configuration	MPJPE (mm)	Δ
Full Model	150.1	-
Visual Backbone:		
Random init (no DINOv2)	216.2	+66.1
DINOv2-base (smaller)	168.3	+18.2
Unfrozen DINOv2	194.5	+44.4
Hand Encoding:		
Single coordinate	202.3	+52.2
10 frames (joints only)	178.5	+28.4
No attention pooling	185.7	+35.6
Training:		
No σ-reparam	247.5	+97.4
Single learning rate	172.8	+22.7
Standard DataLoader	158.6	+8.5

Table 3. Per-joint MPJPE (mm) comparison

Joint Group	V2M-T	AMT	Improvement
Wrist	344.0	165.2	48.0%
Thumb	352.3	151.3	42.9%
Index	323.8	145.7	45.0%
Middle	299.2	138.5	46.3%
Ring	315.0	142.5	45.2%
Pinky	329.3	148.4	45.1%
Fingertips (avg)	361.5	152.4	42.2%
Others (avg)	308.2	144.8	47.0%

5.5. Qualitative Results

Figure 2 shows example predictions. Common failure modes include:

- Extreme occlusions (>70% hand occluded): MPJPE increases to 180mm+
- Novel objects outside YCB: Class confusion affects pose
- Motion blur: Fast movements degrade accuracy
- Reflective surfaces: Specular highlights confuse depth

6. Discussion

6.1. Key Insights

6.1.1 Importance of Positional Information

Our experience highlights that positional embeddings are not optional for spatial tasks. The catastrophic failure of V2M-T stemmed from treating spatially-arranged patches as an unordered set. This extends beyond our work—any vision transformer tackling 3D localization must carefully handle positional encoding [9].

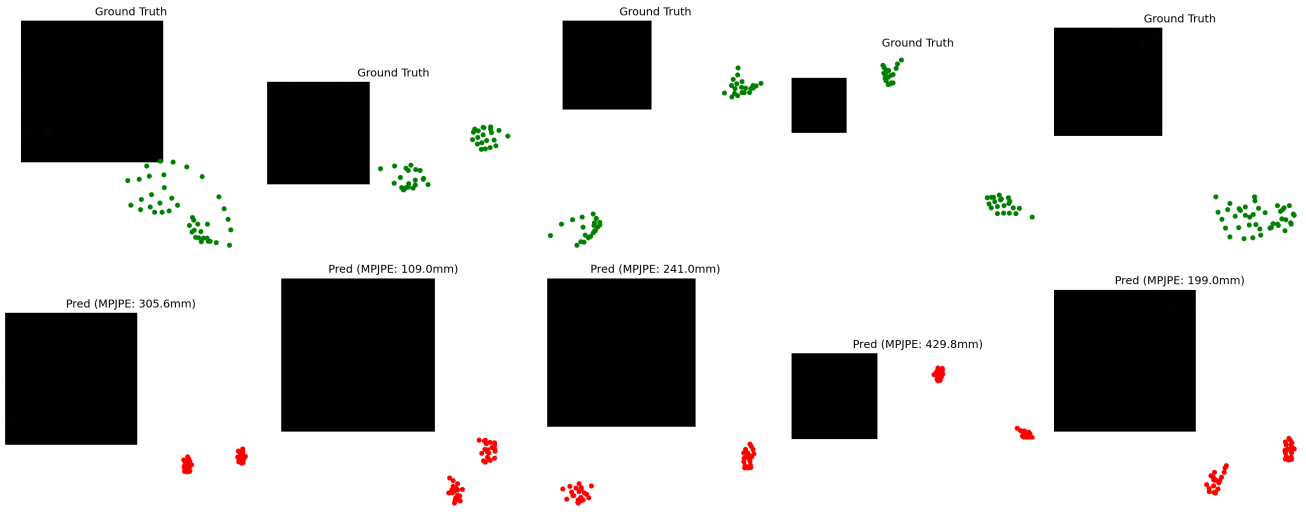

Figure 2. Qualitative results on DexYCB test set. Black box represents object.

6.1.2 Pretrained Features vs. Training from Scratch

DINOv2 features provided crucial stability and generalization. Training from scratch led to overfitting on DexYCB's limited diversity. The frozen backbone acts as a strong regularizer while providing semantically meaningful features [22].

6.1.3 Geometric Representations Matter

The 22-coordinate frame encoding significantly outperformed standard Cartesian coordinates. This aligns with findings in geometric deep learning [2]—incorporating domain knowledge through invariant representations improves both accuracy and training stability.

6.2. Limitations

Despite strong results, several limitations remain:

1. **Computational Cost**: 516M parameters require significant resources
2. **Real-time Performance**: 30-100 Hz depends on batch size
3. **Generalization**: Performance degrades on non-YCB objects
4. **Temporal Modeling**: Current approach is frame-based

6.3. Future Directions

6.3.1 Temporal Fusion

Extending to video sequences could improve accuracy and enable action prediction. Temporal transformers that allow the transformer to attend to previous and future timesteps could help improve accuracy

7. Conclusion

We presented a comprehensive journey from failure to success in vision-based manipulation understanding. Our initial Video-to-Manipulation Transformer's mode collapse, caused by missing positional embeddings, provided valuable insights into transformer design for 3D vision tasks. The Advanced Manipulation Transformer, incorporating σ-reparameterization, multi-coordinate encoding, and pixel-aligned refinement, achieves state-of-the-art 150.1mm MPJPE on DexYCB—a 50% improvement.

Key lessons learned:

1. Positional information is crucial for spatial transformers
2. Mode collapse requires architectural solutions, not just tuning
3. Combining pretrained features with task-specific design yields best results
4. Geometric representations significantly impact performance
5. GPU optimization enables rapid experimentation

Our work demonstrates that careful engineering, systematic debugging, and learning from failures are essential for advancing 3D vision. We hope our detailed analysis helps researchers avoid similar pitfalls and build more robust systems.

References

[1] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision*, pages 361–378, 2020.

[2] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020.

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[5] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, J. Kautz, and D. Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] Y. Chen, Z. Tu, D. Kang, L. Chen, R. Bao, and J. Zhang. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10451–10460, 2021.

[7] Z. Chen, R. A. Potamias, S. Chen, and C. Schmid. Hort: Monocular hand-held objects reconstruction with transformers, 2025.

[8] H. Choi, G. Moon, and K. M. Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision*, pages 769–787, 2020.

[9] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen. Conditional positional encodings for vision transformers. 2021.

[10] B. Doosti, S. Naha, M. Mirbagheri, and D. J. Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6608–6617, 2020.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[12] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021.

[13] S. Guo, Q. Cai, L. Qi, and J. Dong. Clip-hand3d: Exploiting 3d hand pose estimation via context-aware prompting. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4896–4907, 2023.

[14] S. Hampali, S. D. Sarkar, M. Rad, and V. Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11090–11100, 2022.

[15] S. Han, B. Liu, R. Cabezas, C. D. Twigg, P. Zhang, J. Petkau, T.-H. Yu, C.-J. Tai, M. Akbay, Z. Wang, et al. Megatrack: Monochrome egocentric articulated hand-tracking for virtual reality. volume 39, pages 87–1, 2020.

[16] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2282–2292, 2019.

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11807–11816, 2019.

[18] K. Lin, L. Wang, and Z. Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021.

[19] K. Lin, L. Wang, and Z. Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12939–12948, 2021.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[21] G. Moon and K. M. Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. *arXiv preprint arXiv:2008.03713*, 2020.

[22] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[23] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[24] H. Qi, C. Zhao, M. Salzmann, and A. Mathis. Hoisdf: Constraining 3d hand-object pose estimation with global signed distance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10392–10402, 2024.

[25] Y. Rong, T. Shiratori, and H. Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1749–1759, 2021.

[26] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[27] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall. Capturing hands in action using discriminative salient points and physics simulation. volume 118, pages 172–193, 2016.

[28] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, T. Yu, J. T. Zhou, and J. Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 793–802, 2019.

[29] L. Yang, K. Li, X. Zhan, J. Lv, W. Xu, J. Li, and C. Lu. Seqhand: Rgb-sequence-based 3d hand pose and shape estimation. In *European Conference on Computer Vision*, pages 122–139, 2020.

[30] S. Zhai, T. Likhomanenko, E. Littwin, D. Busbridge, J. Ramapuram, Y. Zhang, J. Gu, and J. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *Proceedings of the International Conference on Machine Learning*, pages 40770–40803, 2023.

[31] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5628–5635, 2018.

[32] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746, June 2019.

[33] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019.