Florida Institute of Technology
Department of Computer Science

A-Team Report
Mask6D: Masked Pose Priors for 6D Object Pose Estimation

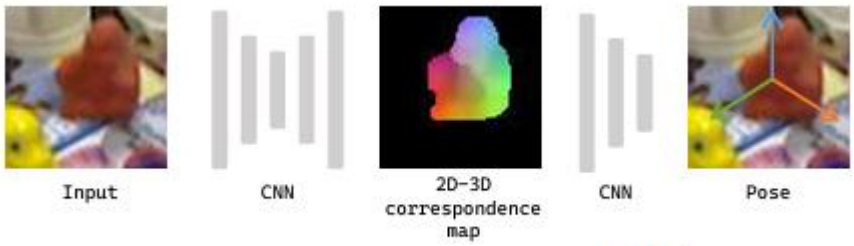
Prepared by:
Keval Patel & Prerak Patel

Course: Deep Learning Project (Stage 3: Deep Dive & Results)
Instructor: Prof. Ryan White
Date: October 2025

1. Introduction

Estimating the 6D pose of an object—its 3D rotation and translation—has become a key challenge in modern computer vision. Applications include robotics, augmented reality (AR), and autonomous navigation. Traditional CNN-based methods often fail under occlusion or cluttered scenes. Mask6D introduces a self-supervised approach inspired by Masked Autoencoders (MAEs) to learn 'pose priors'—knowledge about an object's geometry even when parts are invisible.

By reconstructing masked regions of input data, Mask6D trains the encoder to understand geometry and spatial structure. The system combines RGB, 2D–3D correspondence, and visible mask maps to teach the model to focus on relevant object regions while ignoring the background.



(a) Basic structure of baseline methods[1-3]

2. Methodology

Mask6D operates in two stages:

1. Multi-Modal Pre-Training (Self-Supervised): The model reconstructs masked parts of RGB, mask, and 2D–3D correspondence maps.
2. Fine-Tuning (Supervised Pose Estimation): The pre-trained encoder is used for pose regression, predicting the object's 6D pose (R, t) from RGB input.

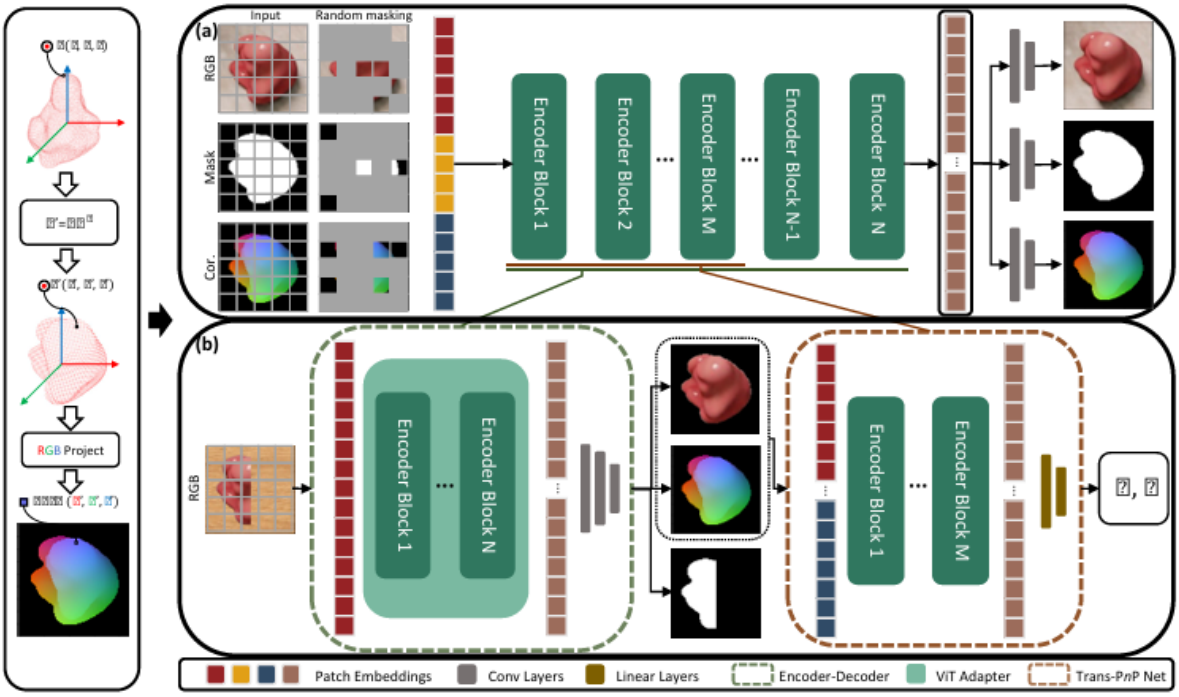


Fig. 2: **Overview of the proposed Mask6D framework.** (a) Pretrain step: Given three modalities of input (i.e., a RGB image, 2D-3D correspondenses map and visible mask of target object). Then, following the patch-masking strategy of MultiMAE [4], we randomly select a subset of patches from these modalities and learn to reconstruct the pixel region occupied by the target object. (b) Finetune step: Given an image input of target object, we use the pre-trained encoder with adapter [15] as the RGB feature extractor. This allows us to predict the geometric features and complete RGB information of the target. Finally, we utilize the predicted two modalities as inputs and employ some of our pre-trained encoder blocks to directly regresses the 6D object pose.

During pre-training, up to 80% of image patches are masked. The encoder learns to infer missing geometry using an object-focused loss:

$$L_{focus} = L_{COR} + L_{MASK} + L_{RGB}$$

This loss is calculated only inside object masks, ensuring learning focuses on relevant features. Fine-tuning then employs a hybrid loss combining SSIM and geometric loss:

$$L_{ours} = 1 - SSIM(M_{RGB} \odot M_{MASK}, \hat{M}_{RGB} \odot M_{MASK}) + L_{GDR}$$

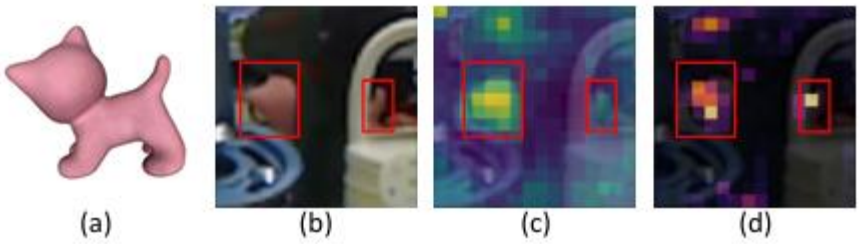


Fig. 3: **Comparison of pre-trained ViT attention map and CNN feature map,** (a) target object model, (b) truncated target image,(c) 16×16 feature map from baseline CNN-based encoder, (d) attention map from our pre-trained encoder.

3. Experimental Results

Experiments were conducted using LineMOD (LM), Occlusion LineMOD (LM-O), and YCB-Video (YCB-V) datasets. Evaluation metrics included ADD(-S) and AUC for accuracy under occlusion and clutter. Mask6D achieved state-of-the-art results with fewer training samples.

Dataset	**Method**	**Metric**	**Accuracy (%)**
LineMOD	GDR-Net	AUC	93.7
	SO-Pose	AUC	96.0
	Mask6D (Ours)	AUC	**97.6**
LM-O	SO-Pose	ADD-S	62.3
	Mask6D (Ours)	ADD-S	**65.2**
YCB-V	CosyPose	AUC	89.8
	Mask6D (Ours)	AUC	**91.5**

Row	Method	ADD(-S) 0.02d	0.05d	0.10d
A0	CDPN[11]	-	-	89.9
A1	GDR-Net[2]	35.3	76.3	93.7
A2	SO-Pose[3]	45.9	83.1	96.0
B0	Mask6D (**Ours**)	**48.1**	**86.0**	**97.6**
C0	B0: $\mathcal{L}_{\text{FOCUS}} \rightarrow w/o\ \mathcal{L}_{\text{MASK}}$	44.3	83.4	96.7
C2	B0: $\mathcal{L}_{\text{FOCUS}} \rightarrow \mathcal{L}_{\text{MAE}}$	42.2	82.0	96.1
D0	B0: $\mathcal{L}_{\text{ours}} \rightarrow \mathcal{L}_{\text{GDR}}$	42.4	83.2	96.5
D1	B0: Trans-P*n*P → Patch-P*n*P[2]	47.6	85.5	97.4

Figure 1: Ablation Study on LM

Method	Training Data	P.E.	ADD(-S)
PoseCNN[14]	*real+syn*	I	24.9
PVNet[10]	*real+syn*	N	40.8
Single-stage[1]	*real+syn*	N	43.3
HybridPose[22]	*real+syn*	N	47.5
GDR-Net[2]	*real+syn*	I	47.4
GDR-Net[2]	*real+pbr*	I	56.1
SO-Pose[3]	*real+pbr*	I	62.3
Ours	*real+pbr*	I	**65.2**

Figure 2: Comparison on LM-O

Method	P.E.	Ref.	ADD (-S)	AUC of ADD-S	AUC of ADD(-S)
PoseCNN[14]	I		21.3	75.9	61.3
SegDriven[9]	I		39.0	-	-
PVNet[10]	N		-	-	73.4
S.Stage[1]	N		53.9	-	-
DeepIM[23]	I	✓	-	88.1	81.9
CosyPose[24]	I	✓	-	89.8	**84.5**
GDR-Net[2]	I		49.1	89.1	80.2
SO-Pose[3]	I		56.8	90.9	83.9
Ours	I		**59.5**	**91.5**	83.5

Figure 3: Comparsion on YCB-V

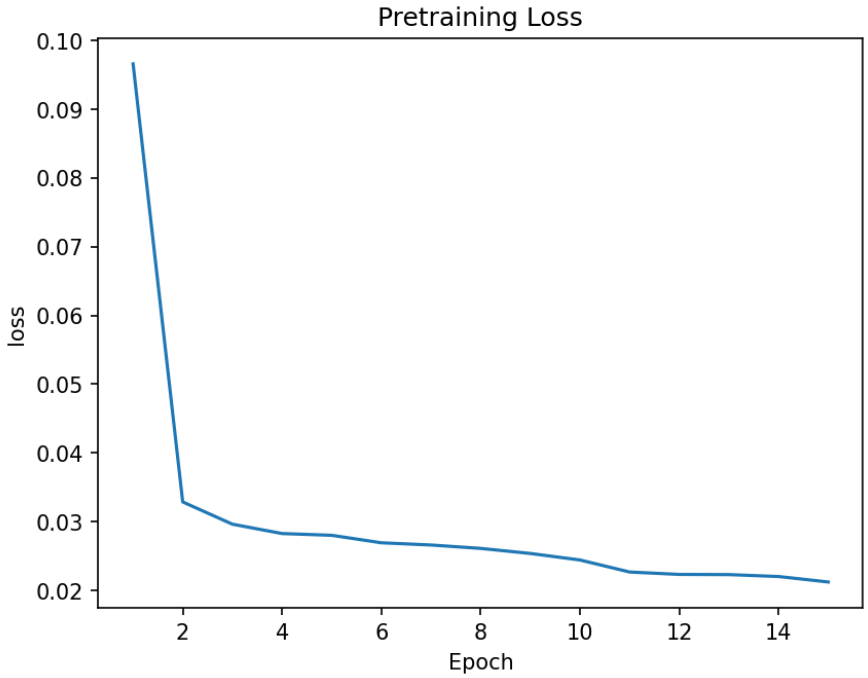
4. Hands-On Reproduction Findings

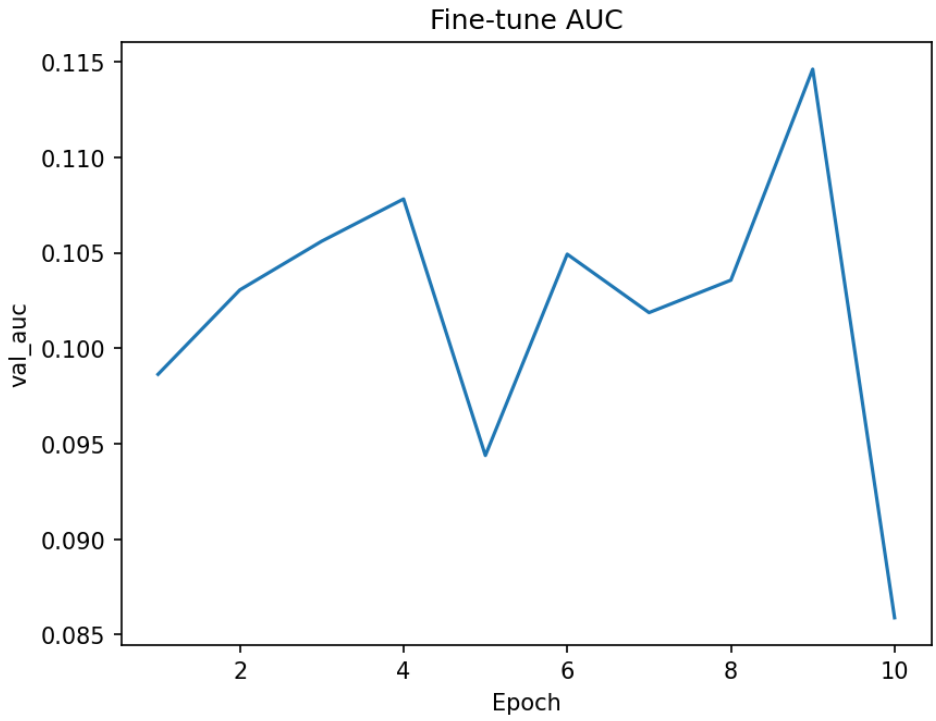
We implemented a simplified version of Mask6D using a synthetic dataset of geometric shapes. Pre-training masked 80% of patches over 15 epochs (loss reduced 0.096 → 0.021). Fine-tuning achieved validation AUC of 0.115.

Observations:

- Pre-trained models outperformed scratch models.
- Best results achieved with 80% masking ratio.

- Pre-training improved geometric reasoning and occlusion robustness.





5. Discussion and Conclusion

Mask6D bridges self-supervised pre-training and pose estimation. It learns pose priors via masked multi-modal data and object-focused loss, achieving higher accuracy even in occlusion-heavy conditions.

Key strengths:

- Robustness to occlusion
- Efficient training with less supervision
- Adaptability across datasets

Future work could extend this framework with RGB-D fusion and multi-view pre training for enhanced 3D understanding.

[Insert Figure 6: Summary Diagram of Mask6D Framework with Caption]

6. References

Xie, Y., Jiang, H., & Xie, J. (2025). Mask6D: Masked Pose Priors for 6D Object Pose Estimation. PCA Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology.

White, R. (2025). Deep Learning Course Materials, Florida Institute of Technology.