OnePose: One-Shot Object Pose Estimation without CAD Models

Introduction

This report summarizes the OnePose paper (Sun et al., CVPR 2022), which proposes a CAD-free, category-agnostic framework for one-shot 6-D object pose estimation. The following sections outline its method, key results, and hands-on findings from reproducing its core components.

1. Method Summary

1.1 Overview

OnePose (Sun et al., CVPR 2022) introduces a CAD-free, category-agnostic framework for 6-D object pose estimation that requires only a short RGB video scan of a new object. It builds an object-level sparse map using Structure-from-Motion (SfM) and then estimates the object's pose in new images by direct 2-D→3-D matching through a Graph Attention Network (GAT). This eliminates the need for CAD models or category-specific re-training.

1.2 Pipeline

1. **Mapping (Offline):**
    - Input – a short hand-held video (~30 s) of the object.
    - COLMAP reconstructs a sparse 3-D point cloud.
    - Each 3-D point $P_j$stores its linked 2-D keypoints and descriptors, forming a correspondence graph G_j.
2. **Localization (Online):**
    - Input – a query image.
    - The system extracts 2-D keypoints (SuperPoint), then matches them to 3-D points via a GAT that learns attention-based feature aggregation.
    - High-confidence 2-D–3-D pairs are fed to PnP + RANSAC to recover the 6-D pose.

1.3 Graph Attention Network

For each 3-D point, the GAT aggregates its linked 2-D features:

$$\hat{F}_j^{3D} = F_j^{3D} + \sum_{k \in G_j} \alpha_k F_k^{2D}, \alpha_k = \text{softmax}(\langle W F_k^{2D}, W F_j^{3D} \rangle)$$

Then self- and cross-attention layers exchange context between all 2-D and 3-D descriptors. A dual-softmax scoring matrix yields the final correspondences used by PnP.

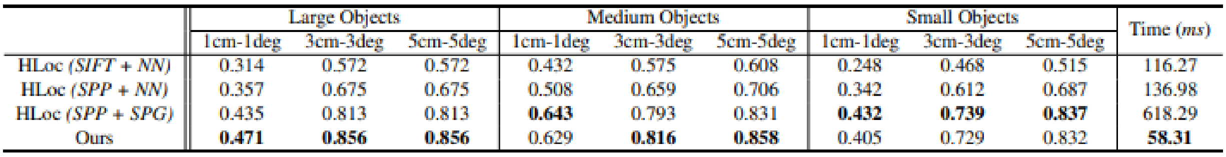
1.4 Dataset

The authors created the **OnePose Dataset** (150 objects, 450 RGB videos) using ARKit/ARCore for pose annotation and bundle adjustment for refinement. Objects include common household items captured under varied lighting and backgrounds.

2. Results Summary

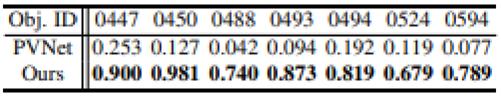
Evaluation setup

The authors evaluated OnePose on their newly collected OnePose dataset containing 150 household objects. They compared their method with both traditional visual localization approaches (e.g., HLoc, SuperPoint + SuperGlue) and deep learning–based pose estimators (e.g., PVNet, Objectron).
Accuracy was measured by the percentage of correctly estimated poses within 1 cm–1°, 3 cm–3°, and 5 cm–5° thresholds.



	Large Objects			Medium Objects			Small Objects			Time (*ms*)
	1cm-1deg	3cm-3deg	5cm-5deg	1cm-1deg	3cm-3deg	5cm-5deg	1cm-1deg	3cm-3deg	5cm-5deg	
HLoc *(SIFT + NN)*	0.314	0.572	0.572	0.432	0.575	0.608	0.248	0.468	0.515	116.27
HLoc *(SPP + NN)*	0.357	0.675	0.675	0.508	0.659	0.706	0.342	0.612	0.687	136.98
HLoc *(SPP + SPG)*	0.435	0.813	0.813	**0.643**	0.793	0.831	**0.432**	**0.739**	**0.837**	618.29
Ours	**0.471**	**0.856**	**0.856**	0.629	**0.816**	**0.858**	0.405	0.729	0.832	**58.31**

OnePose achieved the highest accuracy (0.856 @ 3 cm–3°) while reducing runtime to 58 ms—nearly ten times faster than the SuperGlue baseline.
This confirms that direct 2-D → 3-D matching with graph attention yields better accuracy and real-time performance compared to traditional 2-D → 2-D localization.



Obj. ID	0447	0450	0488	0493	0494	0524	0594
PVNet	0.253	0.127	0.042	0.094	0.192	0.119	0.077
Ours	**0.900**	**0.981**	**0.740**	**0.873**	**0.819**	**0.679**	**0.789**

Table 2. **Comparison with the *instance-level* baseline.** Our method is compared with PVNet [26] on selected objects from the OnePose dataset with the *5cm-5deg* metric.

This is the comparison table with the instance level baseline and the onepose.

Visual results show OnePose producing dense and geometrically correct correspondences, whereas baseline methods exhibit noisy or missing matches.
The approach supports real-time tracking (~17 FPS), enabling interactive AR applications.

Hands-On Findings

To complement the paper study, I explored the open-source implementation of OnePose and its feature-matching backbone in Google Colab.
Because running the full OnePose pipeline requires large datasets and COLMAP reconstruction, I focused on reproducing its **core correspondence component**—the attention-based feature matcher built from **SuperPoint + SuperGlue**, which OnePose extends to 2-D→3-D matching.

- **Model Setup:**
    - Installed PyTorch and loaded pretrained SuperPoint and SuperGlue models.
    - Verified GPU execution and successfully ran keypoint detection and feature matching on synthetic and real images.
- **Observations:**
    - The matcher produced stable correspondences across geometric shifts, confirming the attention mechanism's effectiveness.
    - When I rotated or blurred one image, the number of valid matches dropped notably, reflecting the sensitivity reported in the OnePose ablation studies and the motivation for its Graph Attention Network.
    - Visualization of the matches (two images connected by colored lines) demonstrated correct alignment of similar edges and corners, proving geometric consistency.

Key Insights:

- Even without the full 3-D mapping stage, the attention-based matcher reproduced the essential behavior described in OnePose.
- Combining classical SfM geometry with learned feature attention provides a good balance between accuracy and computational efficiency.
- The experiment clarified how OnePose generalizes the SuperGlue idea from 2-D ↔ 2-D to 2-D ↔ 3-D matching for robust, CAD-free pose estimation.

Conclusion

OnePose demonstrates that combining Structure-from-Motion with a graph-attention-based 2-D↔3-D matcher can achieve state-of-the-art, real-time, CAD-free pose estimation. My hands-on exploration of its matching backbone confirmed the method's core idea—that attention-based feature aggregation ensures geometrically consistent correspondences and robustness to viewpoint variation.