Source-Free and Image-Only Unsupervised Domain Adaptation for Category-Level Object Pose Estimation

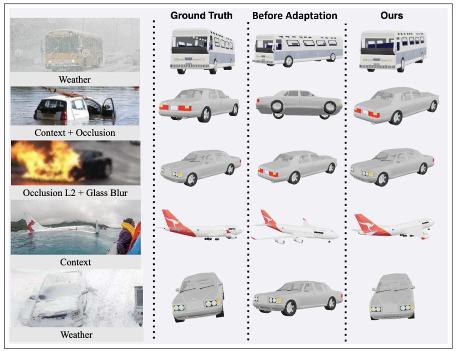
Prakhar Kaushik Aayush Mishra Adam Kortylewski Alan Yuille

Johns Hopkins University | Baltimore, MD, USA

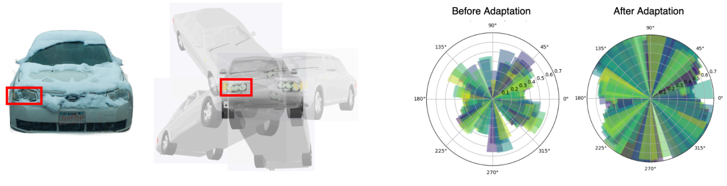
Introduction

We solve the problem of image only and source free unsupervised domain adaptation for category level 3D pose estimation.

Previous domain adaptation methods require some 3D data or depth information in a target domain. Our method, **3DUDA,** does away with this requirement allowing unsupervised domain adaptation to object images found in the real world, even when in presence of nuisances and partial occlusion.



Method Intuition



(a) Local Pose Ambiguity (b) Local Part Robustness

Our method utilizes two key observations

(a) **Local Pose Ambiguity**, i.e. the inherent pose ambiguity that occurs when we can only see a part of the object such that the object may be in a number of viable global poses if we are looking at only a part of it. We utilize this ambiguity to update the local neural vertex features which roughly correspond to object parts, even when the global pose of the object may be incorrectly estimated.

(b) **Local Part Robustness** refers to the fact that certain parts (e.g. headlights in a car) are less affected in OOD data, which is verified by the (azimuth) polar histogram representing the percentage of robustly detected vertex features per image in target domain using the source model (*Before Adaptation*). Even before adaptation, there are a few vertices which can be detected robustly and therefore are leveraged by our method to adapt to the target domain as seen by the increased robust vertex ratio *After Adaptation*.

Results

Table 1: Unsupervised 3D Pose Estimation for OOD-CV (Zhao et al., 2023) dataset

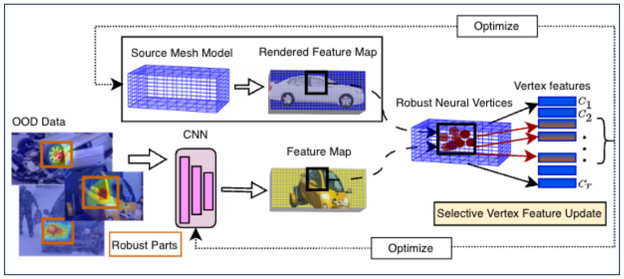
Nuisance	Combined	shape	pose	texture	context	weather
			$\frac{\pi}{6}$ **Accuracy**↑			
Res50-General	51.8	50.5	34.5	61.6	57.8	60.0
NeMo (Wang et al., 2021a)	48.1	49.6	35.5	57.5	50.3	52.3
MaskRCNN (He et al., 2018)	39.4	40.3	18.6	53.3	43.6	47.7
DMNT (Wang et al., 2023)	50.0	51.5	38.0	56.8	52.4	54.5
P3D (Yang et al., 2023)	48.2	52.3	45.8	51.0	54.6	44.5
Ours	**94.0**	**93.7**	**95.1**	**97.0**	**95.5**	**83.1**
			$\frac{\pi}{18}$ **Accuracy**↑			
Res50-General	18.1	15.7	12.6	22.3	15.5	23.4
NeMo (Wang et al., 2021a)	21.7	19.3	7.1	33.6	21.5	30.3
MaskRCNN (He et al., 2018)	15.3	15.6	1.6	24.3	13.8	22.9
DMNT (Wang et al., 2023)	23.6	20.7	12.6	32.6	16.6	33.5
P3D (Yang et al., 2023)	14.8	16.1	12.3	16.6	12.1	16.3
Ours	**87.8**	**82.1**	**69.5**	**92.6**	**89.3**	**90.7**

Table 2: Unsupervised 3D pose estimation results for Occlusion and Extreme UDA setup

(a) **OccL1/L2**: Real Nuisance (OOD-CV (Combined)) + Occlusion (Level1/Level2) (b) **OOD+SN/GB**: Real Nuisance (OOD-CV) + Synthetic Noise (Speckle Noise/Glass Blur) (c) **L1/L2+Spec**: Real Nuisance (OOD-CV) + Occlusion (L1/L2) + Synthetic Noise (Speckle Noise)

	OccL1		**OccL2**		**OOD+SN**		**OOD+GB**		**L1+Spec**		**L2+Spec**	
	$\frac{\pi}{6}$	$\frac{\pi}{18}$	$\frac{\pi}{6}$	$\frac{\pi}{18}$	$\frac{\pi}{6}$	$\frac{\pi}{18}$	$\frac{\pi}{6}$	$\frac{\pi}{18}$	$\frac{\pi}{6}$	$\frac{\pi}{18}$	$\frac{\pi}{6}$	$\frac{\pi}{18}$
NeMo	30.6	10.2	24.1	6.6	32.7	10.2	29.6	9.5	18.6	3.4	15.1	2.7
Ours	**84.6**	**77.1**	**78.7**	**70.4**	**80.5**	**63.0**	**77.7**	**65.9**	**69.4**	**50.4**	**60.6**	**38.9**

Method

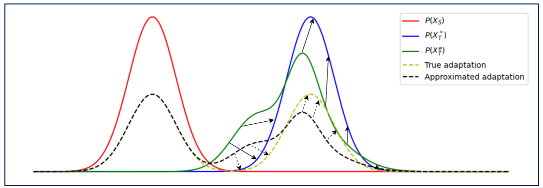


3DUDA Method Overview

(a) We extract neural features from source model CNN backbone $f_i = \phi_w(X_T)$ and render feature maps from the source neural mesh model ($\mathfrak{M}_S$) (using vertex features C_r) and the pose estimate is optimized using feature-level render-and-compare.

(b) For this incorrectly estimated global pose, we measure similarity of every individual visible vertex feature with the corresponding image feature vector in f_i independently and update individual vertex features using average feature vector values for a batch of images.

(c) The neural mesh model is then updated using these changed vertices and the backbone is optimized using the optimized neural mesh.



Theoretical Analysis The elicited target distribution $P(X_T^e)$ found by Selected Vertex Adaptation may not be precisely the same as the true target distribution $P(X_T^*)$, but asymptotically (shown by arrows) it tends to the true distribution and the same happens to the adapted source model.