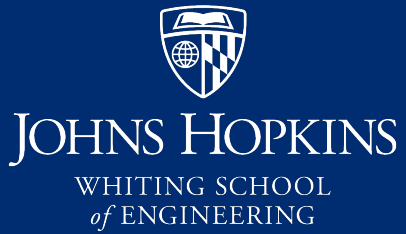
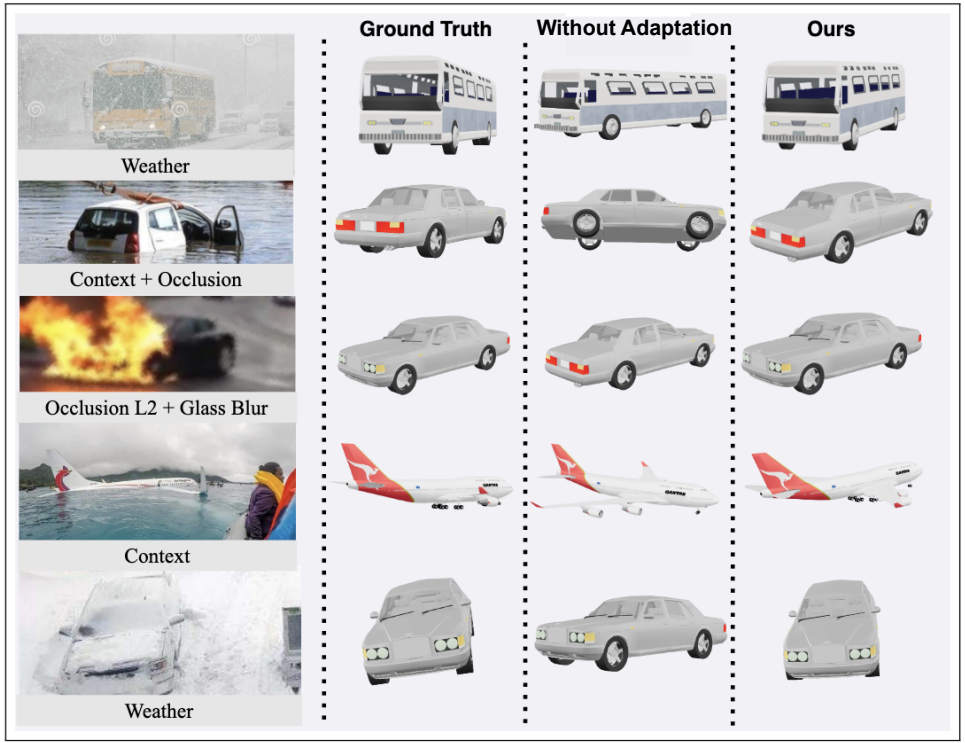


Source-Free and Image-Only Unsupervised Domain Adaptation for Category Level Object Pose Estimation

Prakhar Kaushik Aayush Mishra Adam Kortylewski Alan Yuille

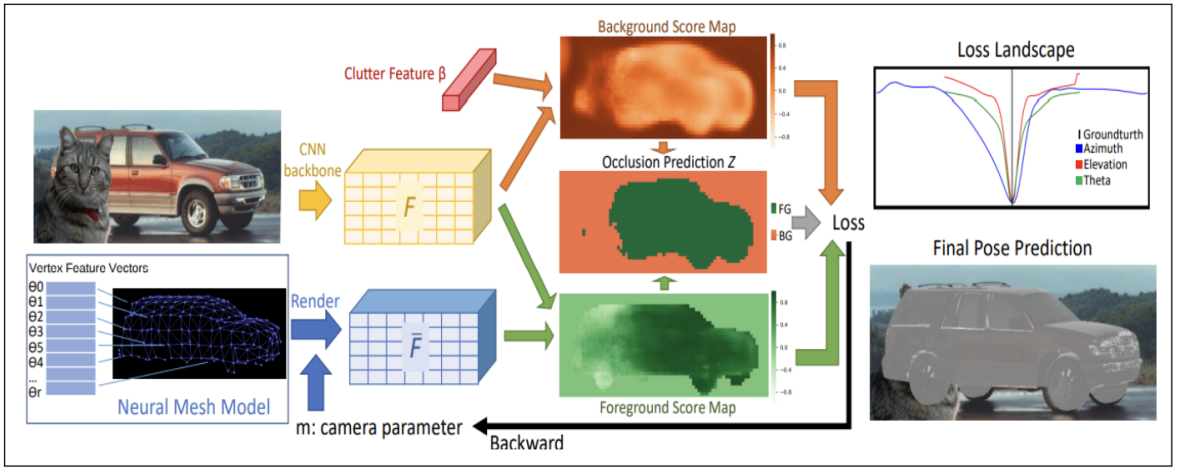
2024 International Conference on Learning Representations

3DUDA: Unsupervised Domain Adaptation for 3D pose estimation



Source-Free and Image-Only Unsupervised Domain Adaptation for Category Level Object Pose Estimation. ICLR 2024

3DUDA: Source Model



Source-Free and Image-Only Unsupervised Domain Adaptation for Category Level Object Pose Estimation. ICLR 2024

3DUDA: Source Model

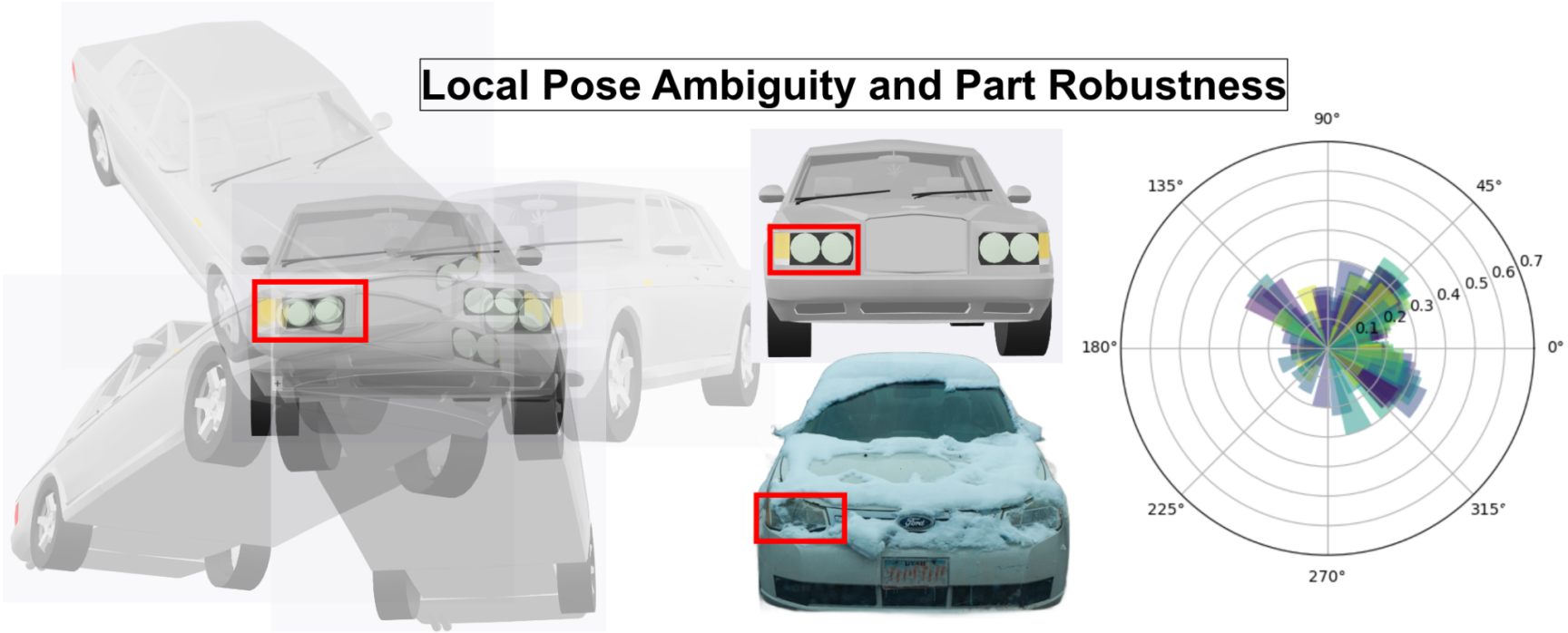
Object Likelihood Computation:

$$p(F|\mathfrak{M}, g, \mathcal{B}) = \prod_{i \in \mathcal{FG}} p(f_i|\mathfrak{M}, g) \prod_{i' \in \mathcal{BG}} p(f_{i'}|B),$$

Object category y,
Source model's neural mesh $\mathfrak{M}_S$ as $\{\mathcal{V}, \mathcal{C}\}$, where $\mathcal{V} = \{V_r \in \mathbb{R}^3\}_{r=1}^{R}$ is the set of vertices of the mesh and $\mathcal{C} = \{C_r \in \mathbb{R}^c\}_{r=1}^{R}$ is the set of learnable features, i.e. neural features.
r denotes the index of the vertices.
R is the total number of vertices.
clutter model $\mathcal{B} = \{\beta_n\}_{n=1}^{N}$ to describe the backgrounds.
Given object pose or camera viewpoint g

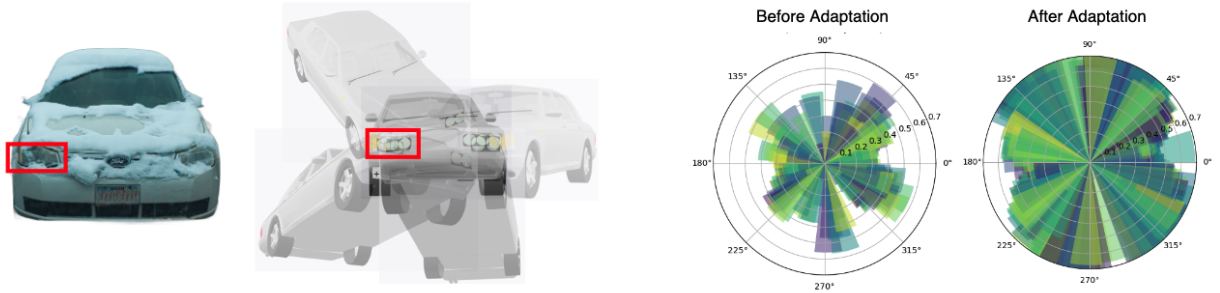
Source-Free and Image-Only Unsupervised Domain Adaptation for Category Level Object Pose Estimation. ICLR 2024

3DUDA: UDA for 3D Pose Estimation



Source-Free and Image-Only Unsupervised Domain Adaptation for Category Level Object Pose Estimation. ICLR 2024

3DUDA: UDA for 3D Pose Estimation



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

Figure 1: 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. We utilize this ambiguity to update the local 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 (OOD-CV (Zhao et al., 2023)) 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*.

3DUDA: Selective Vertex Update

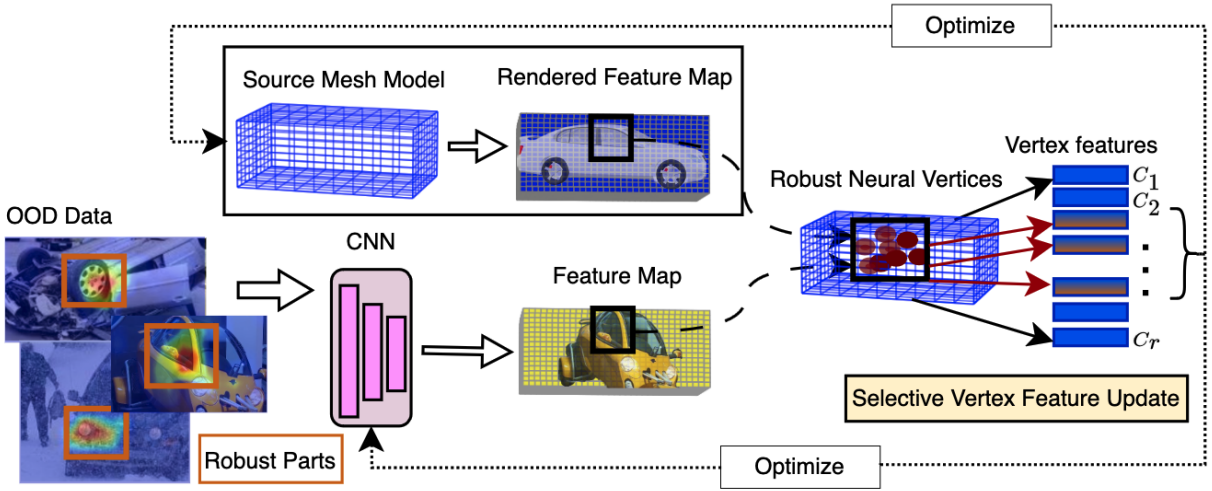
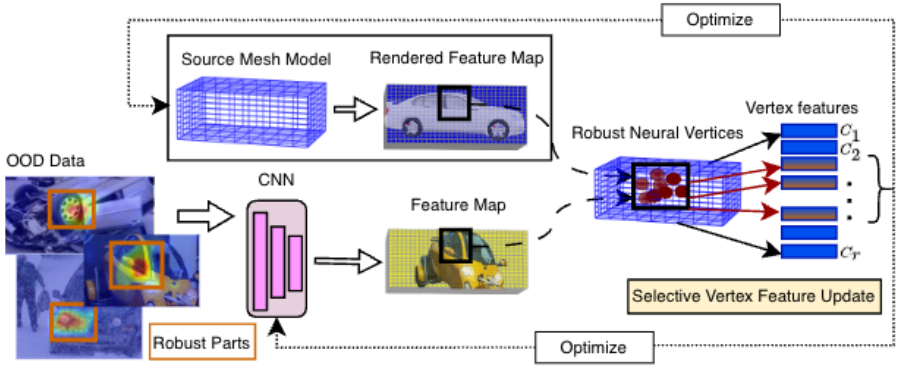


Figure 2: Overview of Our Method (3DUDA)

(a) We extract neural features from source model CNN backbone $f_i = \phi_w(\mathcal{X}_\mathcal{T})$ and render feature maps from the source mesh model ($\mathfrak{M}_\mathcal{S}$) (using vertex features C_r) and the pose estimate is optimized using 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* (Equation 3) and update individual vertex features using average feature vector values for a batch of images (Equation 4). (c) The mesh model is then updated using these changed vertices and the backbone is optimized using the optimized neural mesh.

3DUDA: Selective Vertex Update



$$\mathcal{L}_{sim}(f_{i\to r}, C_r) = Z[\kappa_r] \exp\left(\kappa_r f_{i\to r}^T C_r\right), \qquad \forall i \in \mathcal{FG},\ C_r = \mathfrak{R}(\mathfrak{M}, g)$$

SIREN: Shaping Representations for Detecting Out-of-Distribution Objects. Xuefeng Du, Gabriel Gozum, Yifei Ming, Yixuan Li. NeurIPS 202

$$Z[\kappa_r] = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2}\mathfrak{I}^v_{d/2-1}(\kappa)}$$

(Mardia & Jupp, 2009)

3DUDA: Intuition

Definition 3.1 ***(Vertex K-partition)*** *A vertex K-partition is defined as a partition of the set of vertices (indexed by $r \in \{1, 2, ..., R\}$) into K non-empty mutually disjoint subsets (indexed by $k \in \{1, 2, ..., K\}$). Let the set of vertices in each partitioned subset be denoted by I_k.*

A given vertex K-partition would split the joint distribution P_S into K independent joint distributions (denoted by $P_S^{I_k} = \prod_{r \in I_k} P_S^r$) such that $P_S = \prod_k P_S^{I_k}$. The same extends for the corresponding target domain distributions.

Definition 3.2 ***($k\delta$-subset)*** *For a given sample $\mathcal{X}$ and vertex K-partition, a $k\delta$-subset is defined as $\mathcal{X}^{k\delta} \subseteq \mathcal{X}$ such that, $\mathcal{L}_{sim}(f_{i \to r}, C_r) > \delta_r \; \forall \; r \in I_k$. The corresponding approximation of P^{I_k} under $\mathcal{X}^{k\delta}$ is denoted by $P^{I_k \delta}$.*

Source-Free and Image-Only Unsupervised Domain Adaptation for Category Level Object Pose Estimation. ICLR 2024

3DUDA: Intuition

Assumption 3.3 ***(Piece-wise Support Overlap)*** *There exists a vertex K-partition such that the $k\delta$-subset of the target sample $\mathcal{X}_\mathcal{T}$ satisfies,*

$$|{\mathcal{X}_\mathcal{T}}^{k\delta}| \neq 0 \, \forall k \in \{1, 2, ..., K\}$$

$$\textit{and as } |{\mathcal{X}_\mathcal{T}}^{k\delta}| \to \infty, \prod_k {P_T^*}^{I_k \delta} \to P_T^*$$

This assumption requires the joint distribution of partitioned vertex subsets under $\mathcal{X}^{k\delta}$ to asymptotically approximate the corresponding true distributions. Intuitively, this translates to having enough support in the target domain such that samples satisfying the similarity constraint (Equation 3) in each $k\delta$-subsets approximates the true target distribution of that vertex partition set.

Source-Free and Image-Only Unsupervised Domain Adaptation for Category Level Object Pose Estimation. ICLR 2024

3DUDA: Intuition

Theorem 3.4 *A target domain $\mathcal{X}_\mathcal{T}$ satisfying assumption 3.3, elicits another target domain $\mathcal{X}_\mathcal{T}{}^e$ such that each sample in $\mathcal{X}_\mathcal{T}{}^e$ satisfies the global-pseudo labelling constraint ($\mathcal{L}_{sim}(f_{i\to r}, C_r) > \delta_r \; \forall \, r \in \{1, 2, ..., R\}$). Asymptotically with the size of the domain, $\mathcal{X}_\mathcal{T}{}^e \to \mathcal{X}_\mathcal{T}$.*

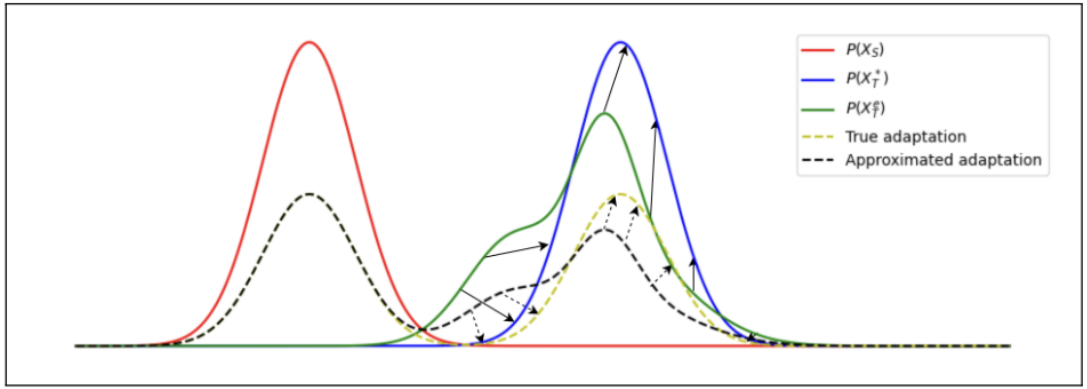


Figure 4: The elicited target distribution $P(X_T^e)$ found by SVA may not be precisely the same as the true target distribution $P(X_T^*)$, but asymptotically (shown by arrows) it tends to the true distribution and the same happens to the adapted source model.

3DUDA: 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**

Source-Free and Image-Only Unsupervised Domain Adaptation for Category Level Object Pose Estimation. ICLR 2024

3DUDA: Results

Table 2: Unsupervised 3D pose estimation results for Pascal3d+ → Corrupted-Pascal3D+ (Metrics : $\pi\backslash 6$ Accuracy ($\frac{\pi}{6}$), $\pi\backslash 18$ Accuracy ($\frac{\pi}{18}$), Median Error (Er))

	$\frac{\pi}{6}\uparrow$	$\frac{\pi}{18}\uparrow$	Er$\downarrow$	$\frac{\pi}{6}\uparrow$	$\frac{\pi}{18}\uparrow$	Er$\downarrow$	$\frac{\pi}{6}\uparrow$	$\frac{\pi}{18}\uparrow$	Er$\downarrow$	$\frac{\pi}{6}\uparrow$	$\frac{\pi}{18}\uparrow$	Er$\downarrow$
	Gaussian Noise			**Shot Noise**			**Impulse Noise**			**Defocus Blur**		
NeMo	43.7	21.3	42.1	50.6	25.3	35.0	45.4	22.2	39.4	72.9	41.8	16.0
Ours	**84.3**	**59.1**	**9.8**	**85.9**	**62.0**	**9.0**	**84.0**	**58.5**	**10.1**	**87.8**	**64.6**	**8.0**
	Glass Blur			**Motion Blur**			**Zoom Blur**			**Snow**		
NeMo	56.7	27.0	33.8	69.7	39.2	18.7	69.0	39.7	19.1	69.9	40.1	18.9
Ours	**86.7**	**62.4**	**8.6**	**88.0**	**63.8**	**8.3**	**87.9**	**65.1**	**8.1**	**87.7**	**64.0**	**8.2**
	Frost			**Fog**			**Contrast**			**Elastic Transform**		
NeMo	73.3	44.1	16.4	85.5	59.0	9.5	74.5	43.8	14.7	77.4	50.3	13.8
Ours	**86.3**	**62.5**	**8.6**	**88.7**	**65.6**	**7.8**	**88.8**	**66.7**	**7.6**	**88.2**	**64.4**	**8.1**
	Pixelate			**Speckle Noise**			**Gaussian Blur**			**Spatter**		
NeMo	77.5	53.0	13.0	67.9	38.3	20.8	68.3	36.5	18.7	72.4	44.2	17.2
Ours	**88.7**	**65.4**	**7.8**	**87.7**	**64.2**	**8.0**	**87.7**	**63.8**	**8.3**	**87.5**	**63.9**	**8.4**

Source-Free and Image-Only Unsupervised Domain Adaptation for Category Level Object Pose Estimation. ICLR 2024

3DUDA: Results

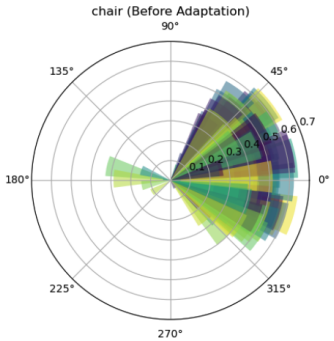
Table 3: Unsupervised 3D pose estimation results for Occlusion and Extreme UDA setup

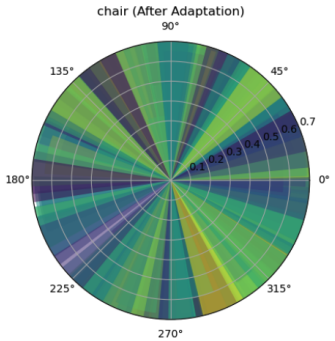
(a) **OccL1/L2**: Real Nuisance (OOD-CV (Combined)) + Occlusion (Level1/Level2) (b) **OOD+SN/GB**: Real Nuisance (OOD-CV) + Synthetic Noise (Speckle Noise/Glass Blur) (c) **L1/L2+Spec**: Real Nuisance (OOD-CV) + Occlusion (L1/L2) + Synthetic Noise (Speckle Noise)

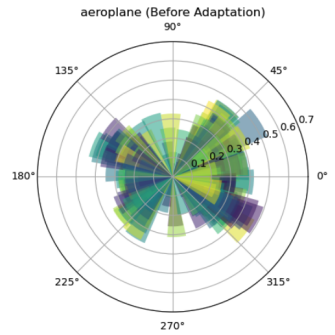
	OccL1		**OccL2**		**OOD+SN**		**OOD+GB**		**L1+Spec**		**L2+Spec**	
	$\frac{\pi}{6}$	$\frac{\pi}{18}$	$\frac{\pi}{6}$	$\frac{\pi}{18}$	$\frac{\pi}{6}$	$\frac{\pi}{18}$	$\frac{\pi}{6}$	$\frac{\pi}{18}$	$\frac{\pi}{6}$	$\frac{\pi}{18}$	$\frac{\pi}{6}$	$\frac{\pi}{18}$
NeMo	30.6	10.2	24.1	6.6	32.7	10.2	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**

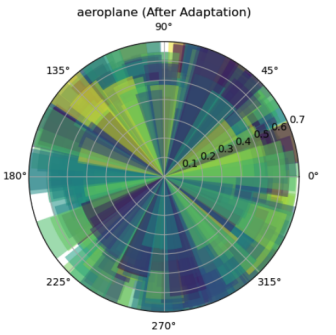
Source-Free and Image-Only Unsupervised Domain Adaptation for Category Level Object Pose Estimation. ICLR 2024

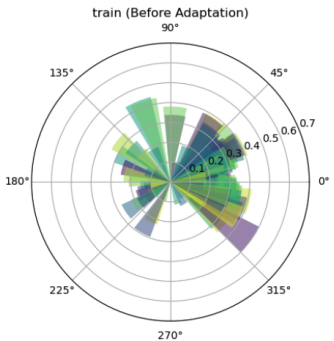
3DUDA: Results

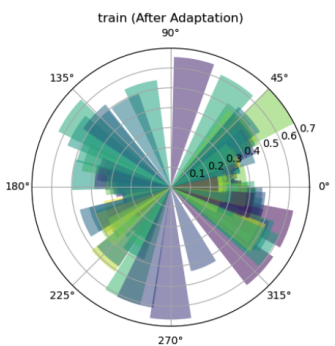


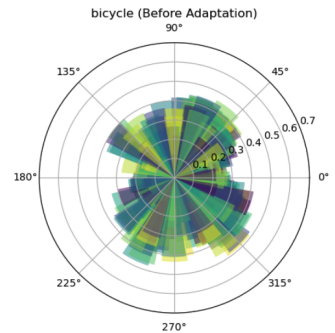


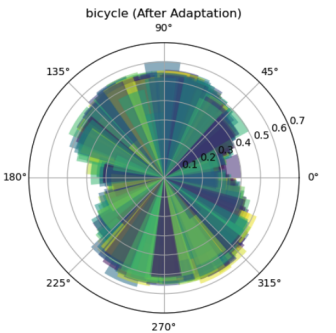












Source-Free and Image-Only Unsupervised Domain Adaptation for Category Level Object Pose Estimation. ICLR 2024