JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

A Bayesian Approach to OOD Robustness in Image Classification

Prakhar Kaushik   Adam Kortylewski   Alan Yuille

Johns Hopkins University | Baltimore, MD, USA

CVPR
JUNE 17-21, 2024
SEATTLE, WA

Introduction

We solve the problem of **Unsupervised Domain Adaptation** to highly corrupted, real world nuisance-ridden (weather, shape, texture, context, 3D pose, etc.) and partially occluded image domains using robust object part representations.

For each object y we learn a generative model $P(F\,|\,y)$ for the feature vectors F. This model is formulated as a mixture model $P(F\,|\,y) = \sum_m P(F\,|\,y, m)$ where the mixture variable m roughly corresponds to the viewpoint of the object. The conditional distributions $P(F\,|\,y, m)$ for the features are factorizable in terms of position so that $P(F\,|\,y, m) = \prod_{a \in \mathscr{D}} P(f_a\,|\,y, m)$, where $a \in \mathscr{D}$ specifies the position in the image. These distributions $P(f_a\,|\,y, m)$ are specified in terms of *von Mises-Fisher (vMF) dictionaries*, with parameters $\Lambda = \{\sigma_k, \mu_k\}$ and by *spatial coefficients* with parameters $\mathscr{A} = \{\alpha^{y,m}_{a,k}\}$.

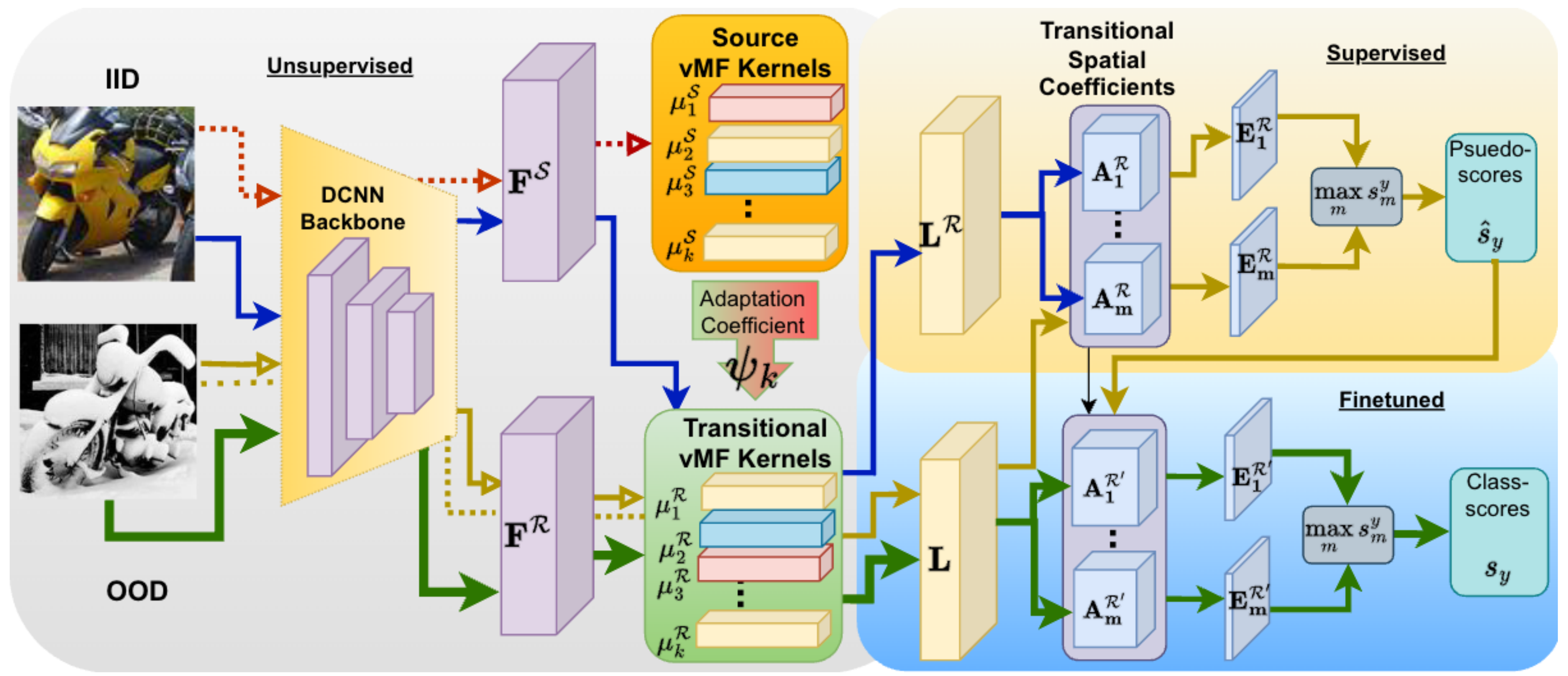
We use the following generative probability distribution for the neural features F conditioned on an object y:

$$P(F|y) = \sum_m P(F|y,m) = \sum_m \prod_{a \in \mathcal{D}} P_a(f_a|y,m)P(m),$$

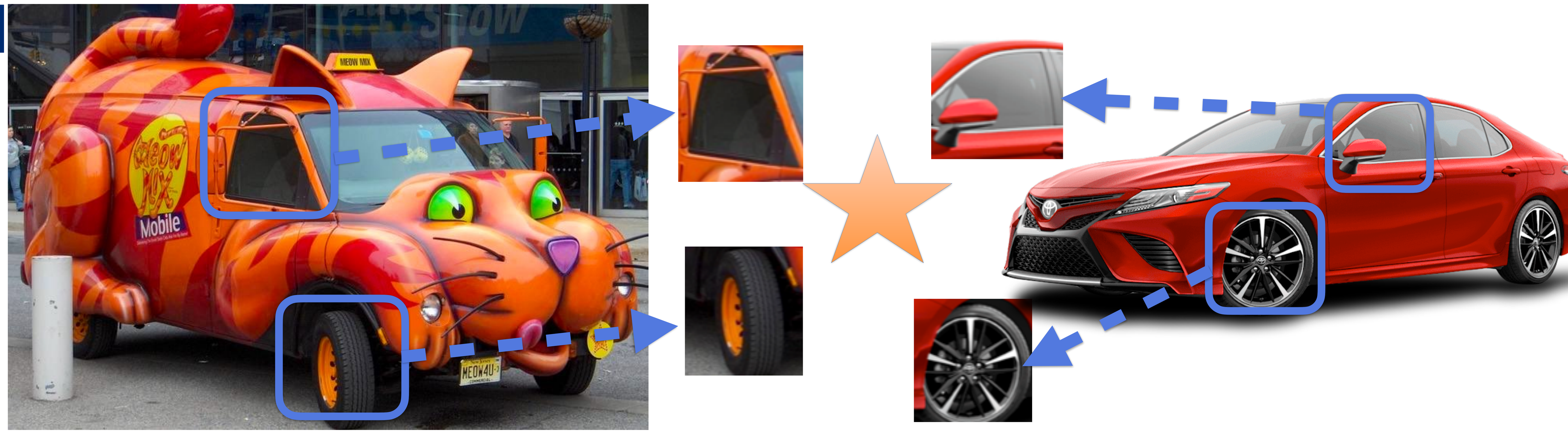
$$P_a(f_a|y,m) = P_a(f_a|\mathcal{A},\Lambda) = \sum_k \alpha^{y,m}_{a,k} P(f_a|\sigma_k,\mu_k),$$

$$P(f|\sigma_k,\mu_k) = \frac{e^{\sigma_k \mu_k^T f}}{Z(\sigma_k)}, ||f|| = 1, ||\mu_k|| = 1,$$

Method



⇢ , ⇢ A DCNN backbone is used to extract the source (IID) $F^{\mathscr{S}}$ and target (OOD) features $F^{\mathscr{R}}$. The source feature vectors $F^{\mathscr{S}}$ are then used to learn the source vMF kernels that are then adapted to the transitional vMF kernels using target domain features $F^{\mathscr{R}}$ and the adaptation coefficients ψ_k in an unsupervised manner. →Transitional Spatial coefficients ($A^{\mathscr{R}}$) are then learned using the transitional vMF likelihood $L^{\mathscr{R}}$ i.e. non-linear activation applied to a convolution of $F^{\mathscr{S}}$ and transitional kernels using source labels. →These spatial coefficients are then finetuned ($A^{\mathscr{R}'}$) using pseudo-scores $\{\hat{s}\}$ generated using the transitional mixture likelihood $E^{\mathscr{R}}$ of target domain features $F^{\mathscr{R}}$. →shows the final feedforward pipeline during inference.



Adaptation by Components

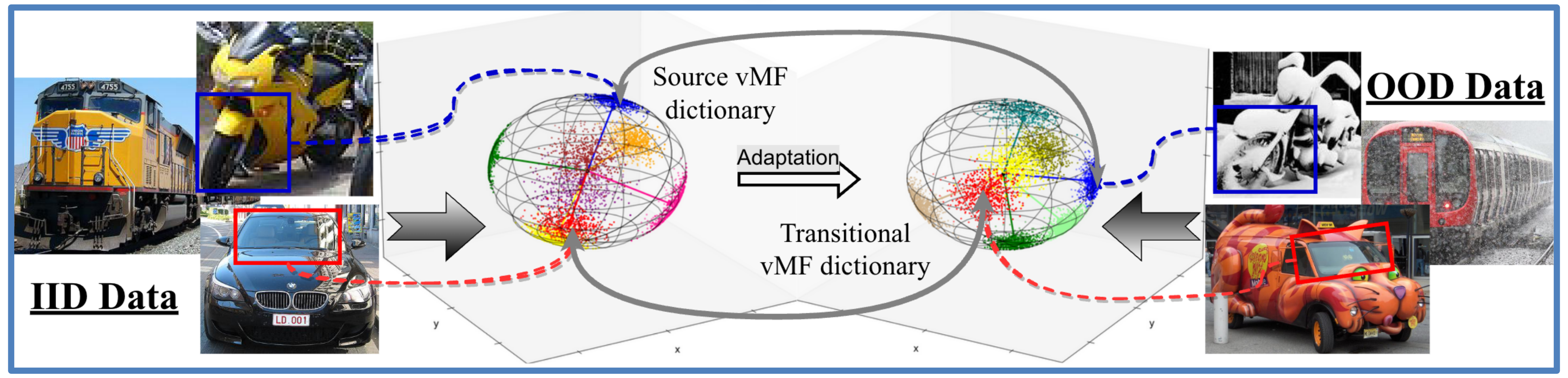
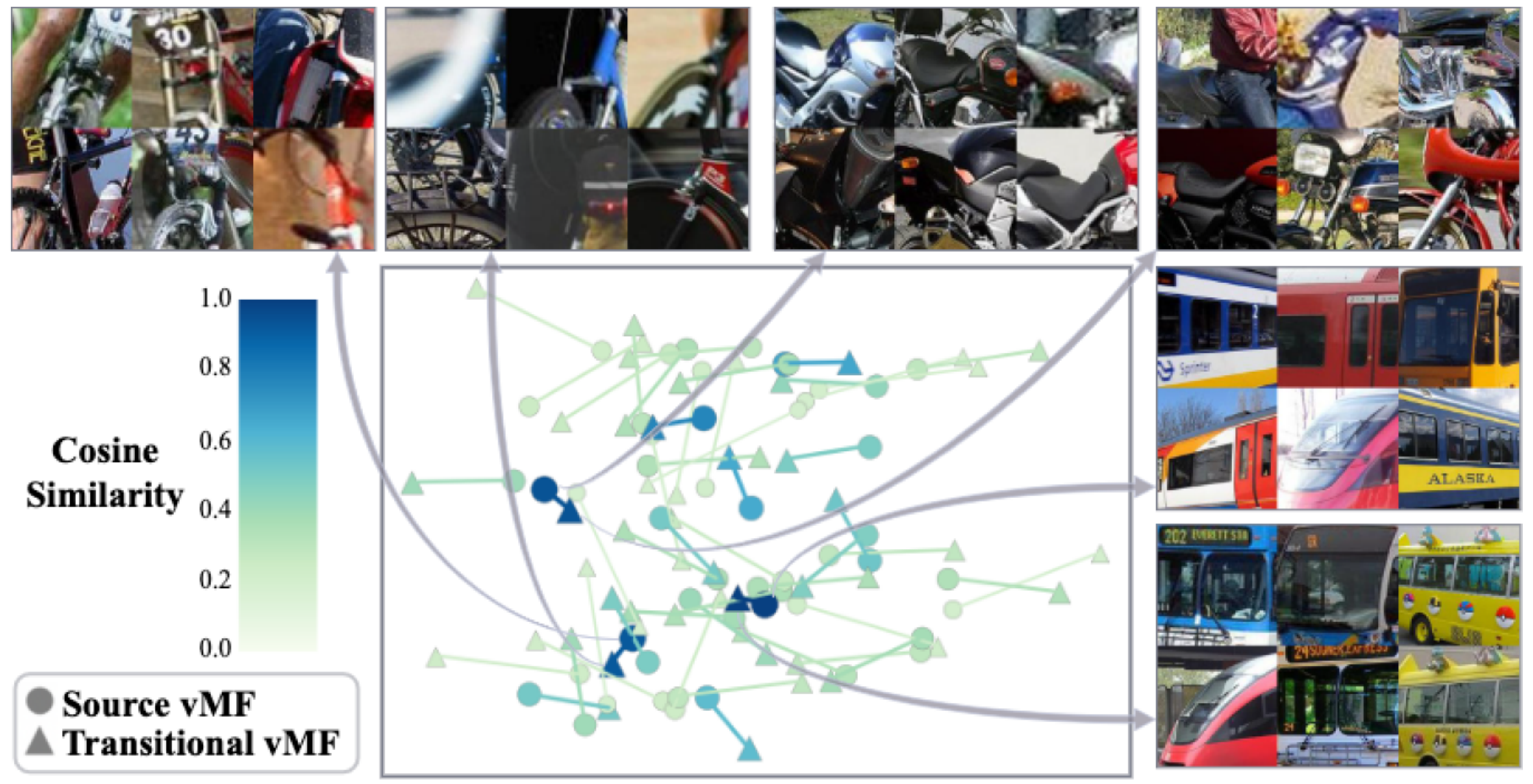


Illustration of the **key principle of adaptation-by-components (alludes to recognition-by-components theory)** underlying our Bayesian approach. We show that clusters of feature vectors learned in an unsupervised manner often resemble **part(component)-like patterns**. We observe that some feature clusters (represented here on a von Mises-Fisher(vMF) manifold) are very similar in both IID and OOD data (illustrated in blue and red boxes), whereas for other feature clusters there is no corresponding equivalent in the other domain. Our Bayesian approach exploits this property by first learning a generative model of feature clusters and their spatial combinations on the IID data and subsequently adapting the model to OOD data via an unsupervised adaptation of the vMF cluster dictionary, while retaining the spatial relations between clusters.

We learn a generative model of image features using vMF distribution mixtures and find parts (representations) of objects in the images which don't change across domain changes. Utilizing these *robust* parts, we adapt to an unlabelled domain in an Expectation-Maximization manner. This technique alludes to the **cognitive science** concepts of **Analysis-by-Components**.

We can do this for **Unsupervised 3D Pose Estimation** too! Check out our **ICLR 2024** work here.

Results



The cosine similarity between source $\Lambda^{\mathscr{S}}$ & transitional vMF dictionary $\Lambda^{\mathscr{R}}$ vectors (represented as circles and triangles) in this conceptual vMF dictionary feature space is represented by the line connecting the circles and triangles. Image patches from the source and target domains roughly corresponding to these vMF dictionary vectors are shown, confirming that some similar image parts are represented by similar vMF dictionary vectors in both domains irrespective of the nuisance factor in the target domain. E.g. (lower right) image patches show windows from different vehicles - parts of objects which do not undergo much change when encountering nuisance factors like change in texture, shape and context of the vehicles.

Table 1. OOD-CV Nuisances Top-1 Classification Results. Occlusion levels greater than 0% represent Occluded-OOD-CV dataset.

Method	Combined				Context				Weather			
Occlusion→	0%	20-40%	40-60%	60-80%	0%	20-40%	40-60%	60-80%	0%	20-40%	40-60%	60-80%
CDAN [25]**	.760	.531	.420	.380	.710	.541	.436	.397	.745	.476	.335	.299
BSP [2]**	.753	.506	.401	.351	.610	.511	.419	.385	.730	.391	.266	.254
MDD [43]**	.780	.551	469	.410	.761	.531	.436	.410	.802	.439	.306	.271
MCD [31]**	.772	.556	.461	.403	.798	.523	.426	.374	.810	.447	.336	.286
MCC [15]**	.785	.582	.492	.434	.730	.577	.454	.420	.767	.503	.376	.362
FixBi [27]**	.821	.534	.478	.399	.802	.542	.445	.409	.755	.489	.358	.335
MIC [13]**	.837	.540	.376	.262	.755	.602	.532	.499	.817	**.612**	.496	.427
ToAlign [40]**	.761	.507	.411	.346	.712	.501	.393	.382	.720	.381	.252	.213
CST [23]**	.840	.579	.539	.477	.687	.491	.452	.411	.813	.558	.397	.356
DUA [26]**	.699	.523	.480	.403	.667	.471	.434	.401	.701	.465	.391	.210
DINE [22]**	.835	.600	.493	.443	.867	.515	.418	.397	.798	.423	.290	.261
RPL [30]	.664	.430	.346	.300	.675	.457	.368	.315	.642	.247	.138	.122
BNA [32]	.653	.426	.343	.298	.580	.397	.342	.278	.635	.295	.179	.171
CompNet [21]	.720	.506	.462	.415	.790	.517	.454	.369	.683	.434	.398	.362
UGT (Ours)	**.850**	**.620**	**.570**	**.501**	**.875**	**.624**	**.565**	**.511**	**.856**	.600	**.528**	**.465**