LATENT ENHANCING AUTOENCODER FOR OCCLUDED IMAGE CLASSIFICATION
SUPPLEMENTARY MATERIAL

Introduction

This document provides supplementary material for the paper titled "Latent Enhancing AutoEncoder for Occluded Image Classification" submitted to the regular track of the ICIP 2024. This document consists of details of the architecture of the LEARN, illustration of improvements in inter-class differentiability in latent space for OccludedPASCAL3D+ dataset (hereafter referred to as Pascal), and detailed classification results. Some of this content was omitted from the manuscript due to brevity, however it will be helpful for readers to comprehend the ideas and results presented in the paper.

Contents

1 Architecture of the LEARN

In this section, we elaborate the layer-wise architecture of the LEARN for enhancing the robustness of the classification CNN (backbone) to occlusions of input images. While most of the works on occlusion handling focus on variants of Compositional Networks, we exploit the inherent capability of AutoEncoders (AEs) to reconstruct the features by imposing class-specific constrains on the latent space of the AE.

Layer	**VGG16**		**ResNet-50**	
	Dimensions (c h w)	**#Parameters**	**Dimensions (c h w)**	**#Parameters**
Conv2d	$64 \times 7 \times 7$	294,976	$64 \times 8 \times 8$	1,179,712
AvgPool2d	$64 \times 5 \times 5$	–	$64 \times 6 \times 6$	–
Conv2d	$64 \times 5 \times 5$	36,928	$64 \times 6 \times 6$	36,928
Conv2d	$32 \times 5 \times 5$	18,464	$32 \times 6 \times 6$	18,464
ConvTranspose2d	$64 \times 5 \times 5$	18,496	$64 \times 6 \times 6$	18,496
ConvTranspose2d	$64 \times 7 \times 7$	102,464	$64 \times 8 \times 8$	102,464
ConvTranspose2d	$512 \times 7 \times 7$	295,424	$2048 \times 8 \times 8$	1,181,696

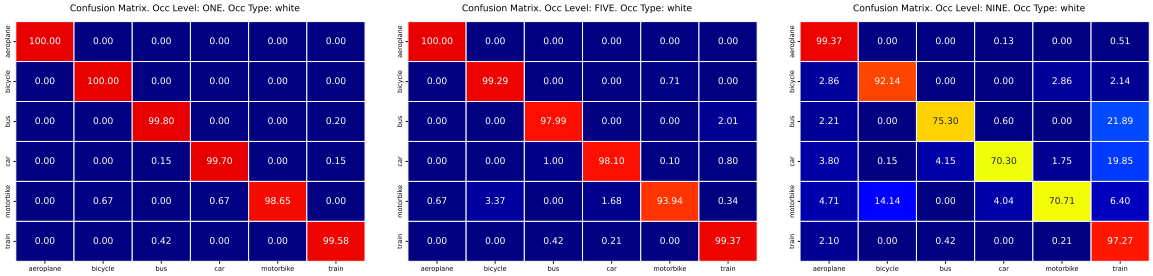
Table 1: Architecture of the LEARN for VGG16 and ResNet-50 backbones.

2 Classification Results on the OccludedPASCAL3D+ dataset

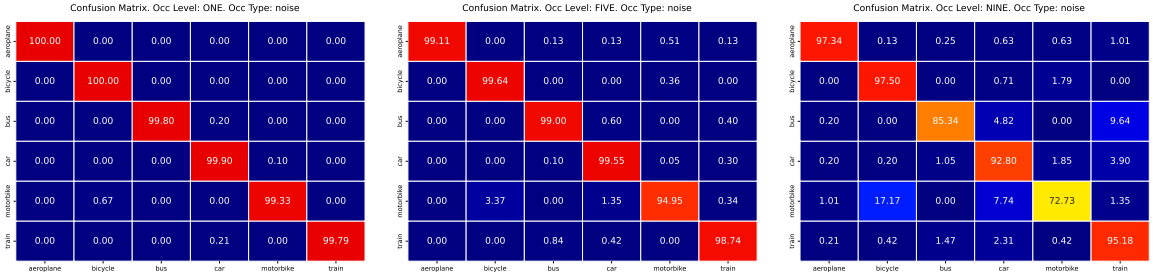
Occ. Area	**L0: 0%**	**L1: 20-40%**				**L2: 40-60%**				**L3: 60-80%**				**Mean**
Occ. Type	-	w	n	t	o	w	n	t	o	w	n	t	o	
baseline	99.9	98.2	97.6	97.9	94.7	94.1	90.6	90.5	72.2	69.8	53.2	50.1	48.1	81.3
CoD*	92.1	92.7	92.3	91.7	92.3	87.4	89.5	88.7	90.6	70.2	80.3	76.9	87.1	87.1
VGG+CoD*	98.3	96.8	95.9	96.2	94.4	91.2	91.8	91.3	91.4	71.6	80.7	77.3	87.2	89.5
TDAPNet*	99.3	98.4	98.6	98.5	97.4	96.1	97.5	96.6	91.6	82.1	88.1	82.7	79.8	92.8
CompNet*	99.3	98.6	98.6	98.8	97.9	**98.4**	98.4	97.8	94.6	**91.7**	90.7	**86.7**	88.4	**95.4**
Proposed	**100**	**99.7**	**99.8**	**99.6**	**99.0**	98.3	**99.0**	**98.1**	**96.1**	80.5	**91.9**	84.4	**89.3**	95.1

Table 2: Performance evaluation of LEARN on the Pascal dataset, using VGG16 backbone, on varying levels and types of occlusions. Occlusion level signifies the percentage of the input image that is occluded and the occlusion types are, w=white noise, n=random noise, t=texture and o=natural objects. For the methods marked with *, we report the results from [1].

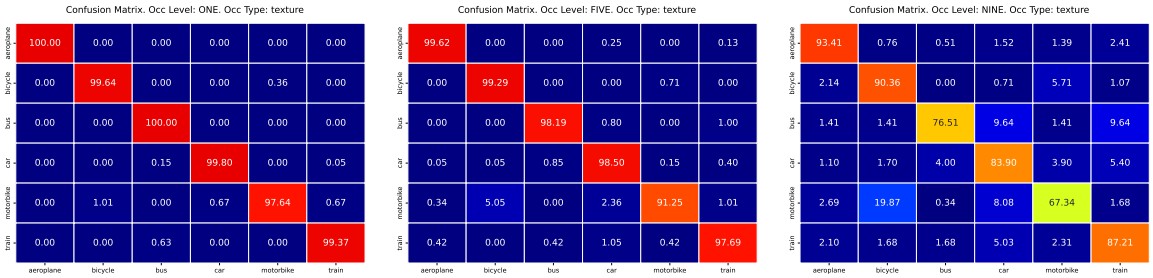
The above table (also a part of the main paper) provides consolidated results of incorporating the LEARN in VGG16 backbone. The classification accuracy of Pascal dataset (test partition) for different levels and types of occlusion are provided. Fig. 1 shows the corresponding confusion matrices for 6 test classes as defined in the experimental protocol.



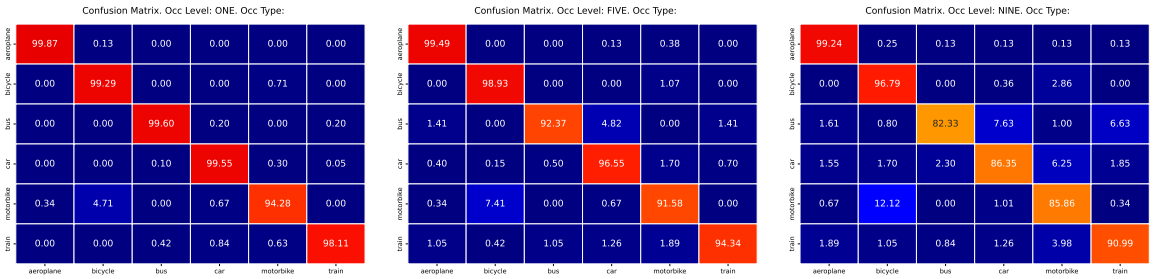
(a) level-ONE-white (b) level-FIVE-white (c) level-NINE-white



(d) level-ONE-noise (e) level-FIVE-noise (f) level-NINE-noise



(g) level-ONE-texture (h) level-FIVE-texture (i) level-NINE-texture

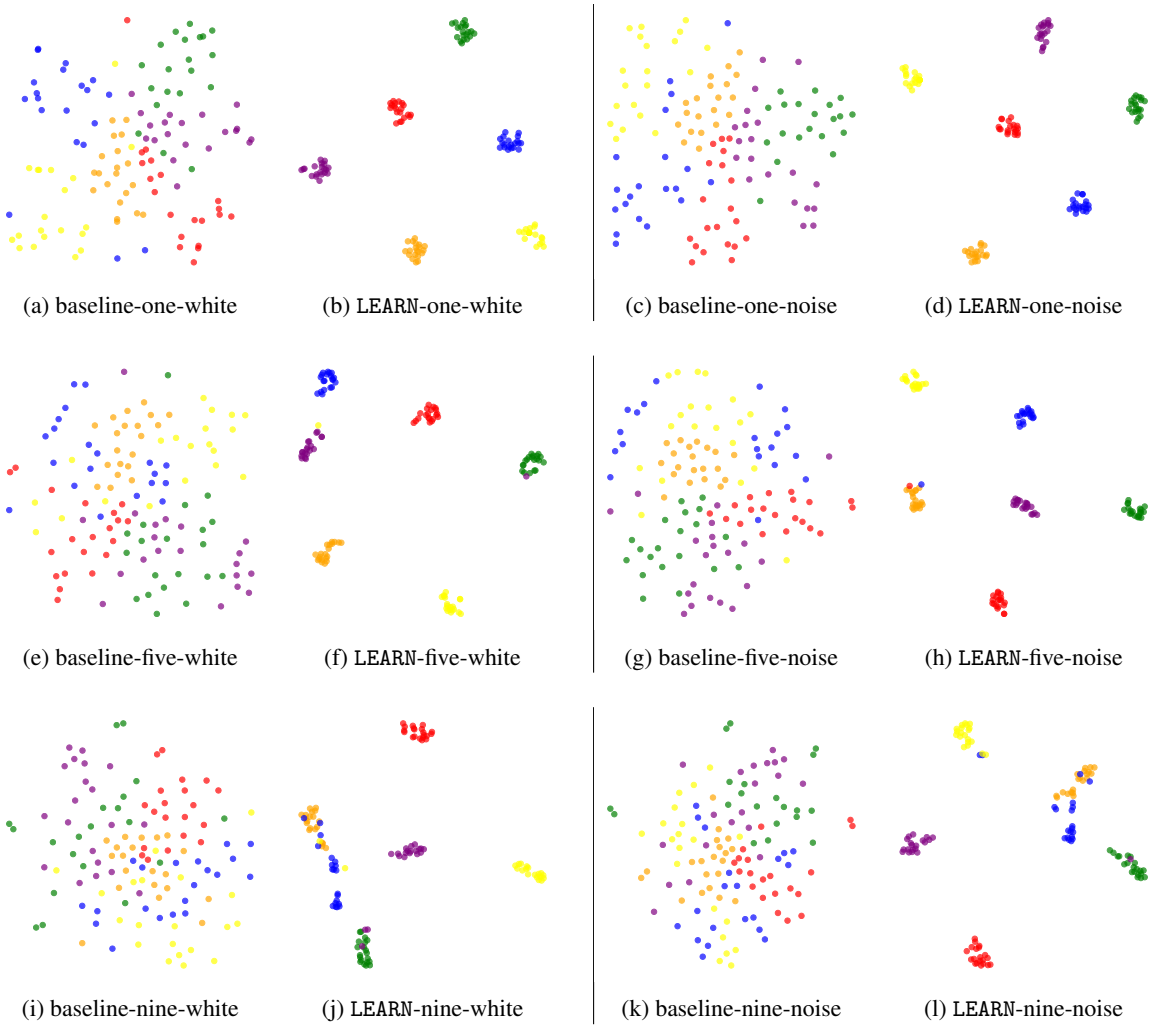


(j) level-ONE-object (k) level-FIVE-object (l) level-NINE-object

Figure 1: Confusion matrices corresponding to classification on the Pascal test dataset by LEARN. Each row represents result on a specific type (white, noise, texture and object) of occlusion, and each column represents result on a specific intensity (level-one, five and nine) of occlusion in test images.

3 Analysis of the latent space of LEARN on the OccludedPASCAL3D+ dataset

The proposed AE based LEARN optimizes the latent space by imposing auxiliary constraints to minimize the intra-class separation while simultaneously maximizing the inter-class separation. This optimization enhances the recovery of intermittent features of the occluded data. Figs. 2 and 3 show all the t-sne plots of the latent spaces created using the baseline and LEARN (using VGG16 as the backbone) on the Pascal dataset. As seen in these plots, LEARN enables significantly better clustering for all types and levels of occlusions, i.e. the features of images of same class are grouped together despite presence of high noise content due to occlusions, and the features of images of different classes are mostly clustered apart from each other. Hence, features are not mis-classified in the presence of occlusions.

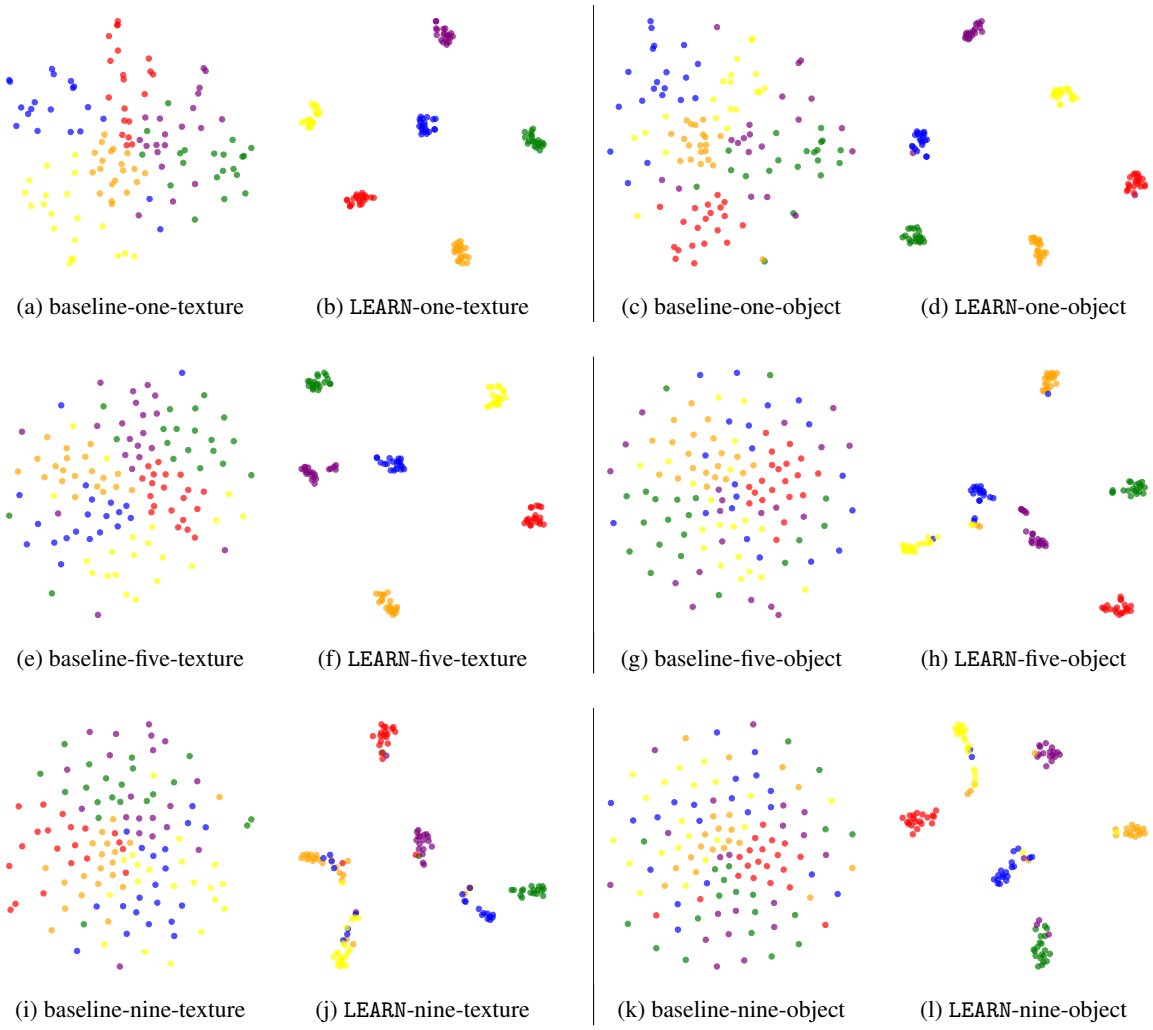


(a) baseline-one-white (b) LEARN-one-white (c) baseline-one-noise (d) LEARN-one-noise

(e) baseline-five-white (f) LEARN-five-white (g) baseline-five-noise (h) LEARN-five-noise

(i) baseline-nine-white (j) LEARN-nine-white (k) baseline-nine-noise (l) LEARN-nine-noise

Figure 2: t-SNE plots of latent space using baseline and LEARN (with VGG16 backone) on the Pascal test dataset. Each row corresponds to a different level of occlusion intensity (one, five and nine) for two types of occlusions– white-patch and noise.



(a) baseline-one-texture (b) LEARN-one-texture (c) baseline-one-object (d) LEARN-one-object

(e) baseline-five-texture (f) LEARN-five-texture (g) baseline-five-object (h) LEARN-five-object

(i) baseline-nine-texture (j) LEARN-nine-texture (k) baseline-nine-object (l) LEARN-nine-object

Figure 3: t-SNE plots of latent space using baseline and LEARN (with VGG16 backone) on the Pascal test dataset. Each row corresponds to a different level of occlusion intensity (one, five and nine) for two types of occlusions– texture and random objects.

References

[1] A. Kortylewski, J. He, Q. Liu, and A. L. Yuille, "Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion," in *IEEE/CVF CVPR*, 2020, pp. 8940–8949. 3