Latent Enhancing AutoEncoder for Occluded Image Classification

Ketan Kotwal
ketan.kotwal@idiap.ch

Tanay Deshmukh
tanaydeshmukh96@gmail.com

Preeti Gopal
preetigopalindia@gmail.com

Abstract

Large occlusions result in a significant decline in image classification accuracy. During inference, diverse types of unseen occlusions introduce out-of-distribution data to the classification model, leading to accuracy dropping as low as 50%. As occlusions encompass spatially connected regions, conventional methods involving feature reconstruction are inadequate for enhancing classification performance. We introduce `LEARN`: **L**atent **E**nhancing fe**A**ture **R**econstruction **N**etwork– An auto-encoder based network that can be incorporated into the classification model before its classifier head without modifying the weights of classification model. In addition to reconstruction and classification losses, training of `LEARN` effectively combines intra- and inter-class losses calculated over its latent space—which lead to improvement in recovering latent space of occluded data, while preserving its class-specific discriminative information. On the OccludedPASCAL3D+ dataset, the proposed `LEARN` outperforms standard classification models (VGG16 and ResNet-50) by a large margin and up to 2% over state-of-the-art methods. In cross-dataset testing, our method improves the average classification accuracy by more than 5% over the state-of-the-art methods. In every experiment, our model consistently maintains excellent accuracy on in-distribution data[1].

1 Introduction

Recent research has shown that despite their high performance in various areas, deep convolutional neural networks (CNNs) encounter difficulties with occluded data. This challenge presents significant obstacles in applications such as autonomous driving [1], video surveillance [2], and medical imaging [3], where precise classification of occluded objects is vital. Misidentifying occluded objects can lead to accidents or decision-making errors, particularly in the context of autonomous driving. Thus, addressing the issue of occlusions is critical for strengthening the robustness and reliability of these models for consistent and trustworthy deployment.

The deterioration in performance of CNNs, (as shown in Fig. 1), can be attributed to the complexity in handling occlusions that occur at different scales, locations, and ratios [4]. Addressing all possible occlusion patterns during training a classification model is extremely demanding due to their vast number and variability [5]. For instance, the occluding object (*occluder*) may be new and unseen in the training data. Additionally, it can exist in any of the innumerable positions possible, and occlude the target object (*occludee*) to any extent (*level*). Consequently, deep learning-based models struggle to generalize well to partially occluded objects. The issue of occluded data can also be viewed from the perspective of out-of-distribution (OOD) data. Occlusions bring about significant visual variation, leading to shifts and deviations in the underlying data distribution compared to the training dataset. Therefore, CNN models trained on specific datasets often face challenges in accurately classifying occluded objects because they have not been exposed to such variations during training.

[1]Supplementary Material can be downloaded from: IEEE SigPort repository

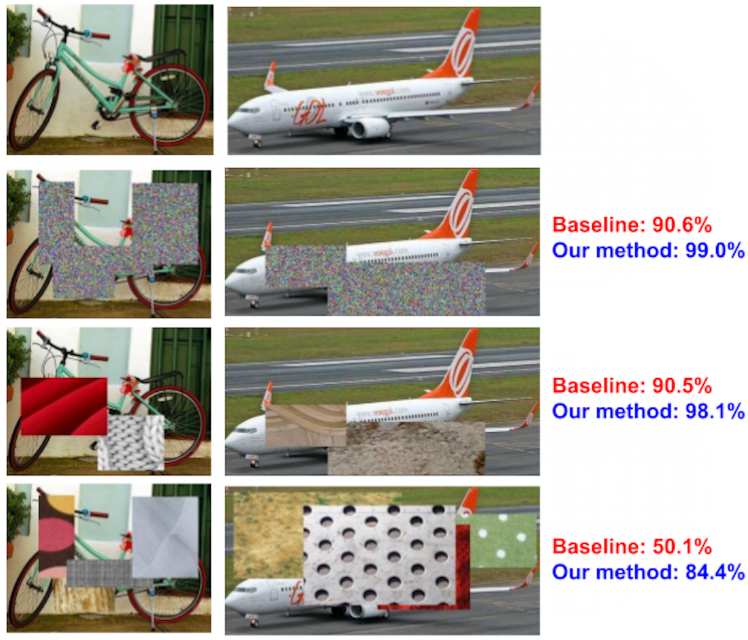


Figure 1: Examples of clean and occluded images from OccludedPASCAL3D+ dataset. The clean images are at the top row, and their occluded versions with patches of noise (of level 5), texture (of level 5) and random objects (of level 9) are shown in subsequent rows respectively.

The improvement towards resilience to OOD data is a significant area of study within deep learning and computer vision. This focus has been recognized through specialized challenges and competitions, such as the OOD-CV challenge series at ICCV 2023[2] and ECCV 2022[3]. Recent benchmark studies [6, 7] categorize the OOD problem based on nuisance factors such as illumination, pose, context, weather variations, and occlusions. These studies indicate that among nuisance factors, occlusion presents the most demanding challenge where classification accuracy for a ResNet-50 architecture [8] model drops by up to 25% on the OOD-CV-v2 benchmark dataset. Consequently addressing occluded data remains pivotal in enhancing robustness and reliability of deep CNN-based classification models.

In this work, we tackle the problem of classifying objects that are subject to various levels and forms of occlusions. The key to accurately classifying an occluded object lies in properly reconstructing the underlying latent space that ascribes the correct class. Moreover, it is crucial that any approach for handling the OOD data has minimal impact on images of non-occluded objects (*i.e.*, should not degrade in-distribution data). This challenge is akin to removing noise from large and connected components of complex class-specific features obtained from deep CNNs. We harness the inherent capability of autoencoders (AE) to acquire robust and compact representations of data. We design an AE-based deep network– `LEARN` (**L**atent **E**nhancing fe**A**ture **R**econstruction **N**etwork)– that not only learns to enhance specific classes' latent features but also reconstructs occluded features. This is accomplished by training `LEARN` on samples containing both occluded and non-occluded (clean) objects from corresponding classes. For this purpose, we create occluded images with various types and levels of occlusion on-the-fly by randomly occluding different parts of the clean images, mimicking real-world scenarios. As occlusions may hinder as much as 70+% of the spatial area (see last two rows of Fig. 1), it is extremely difficult to recover the true (or clean) features solely by observing the local neighborhood. Hence, we incorporate auxiliary loss terms which constrain the similarity between the latent space of occluded and clean images while maintaining sufficient inter-class distances within latent space. Lastly, we utilize classification loss (through frozen fully connected (FC) layers) to supervise `LEARN`'s training and align it with the backbone model. We do not modify any parameters of the backbone model (apart from the final FC layer finetuned to given number of classes), allowing for easy integration of our proposed `LEARN` into inference pipelines with minimal adjustments necessary.

The contributions of our work can be summarized as follows:

[2]https://www.ood-cv.org

[3]https://www.ood-cv.org/2022/index.html

- We design an auto-encoder based network that can be incorporated within a classification CNN to enhance its robustness at classifying objects with varying degrees and types of occlusions. It can seamlessly work across different CNN classification models.
- We propose auxiliary loss functions that facilitate efficient reconstruction of occluded data in the latent space while simultaneously enhancing their discriminative capability.
- Our results demonstrate that the `LEARN` model leads to significant improvements in different types of occluded images, and outperforms recent state-of-the-art methods. It also maintains excellent accuracy for clean, non-occluded images. With only 0.7M parameters, the `LEARN` for VGG16 backbone improves the classification accuracy from 55 % (baseline) to 86% for occlusions as high as 60–80% of object area. For the same set of occlusions, the accuracy of classification is improved by 28% over the baseline and 3% over state-of-the-art methods for a ResNet-50 backbone with 2.5M parameters.

2 Related Work

Although the deep CNN-based classification models are highly discriminative [9, 10], their performance rapidly deteriorates with increasing level and complexity of occlusions. On the other hand, compositional models are more robust [11] since they represent objects as a composition of various spatial components of the image. Hence their overall inference is accurate even when a few spatial components are wildly different due to partial occlusions. In [5], the classical CNN model and a compositional mixture model are stacked together leveraging the advantages of both networks. Here the compositional model is trained with the feature vectors in the final convolutional layer, making the model independent of attributes such as illumination and pose. During inference, the image is first passed through the CNN. Only when this module classifies with low certainty indicating the possible presence of occlusions, the features are fed to the compositional model which predicts the final class. In an extension of this work, Kortylewski *et al*. introduced Compositional Neural nets wherein the final fully connected classification layer is replaced by a differentiable composition of mixture models [12]. The parameters of the mixture model were trained such that for each category of object, one of the mixture models gets activated. The Compositional Neural Net was further modified in [13] to learn separate representations of the object and its background (context). In cases of strong occlusions, this prevents the model from falsely detecting the object based on the features of the context. This strategy combined with an estimation of the object's bounding box led to improved performance on severely occluded vehicle images from MS-COCO [14] and PASCAL3D+ [15] datasets.

Another way of handling occlusions is by introducing better augmentation strategies in order to introduce better generalization to the models. In this regard, the idea of Soft Augmentation was recently introduced in [16], wherein the confidence in target label was reduced non-linearly based on the level of augmentation of the training sample. This led to better object classification under occlusions. However, the reported occlusion types were synthetic consisting of box-shaped constant intensity patches, mimicking the original image being cropped to different degrees. The performance of this strategy on images with realistic occluders is yet to be explored.

Available Datasets: The availability of standard annotated occluded images is limited. Simple occluders usually consist of a constant gray-scale mask, patches of white noise or textures, while realistic occluders are real-life objects themselves. There are limited number of datasets of the latter kind. As an example, in OccludedPASCAL3D+ dataset (hereafter referred to as Pascal) [13], occlusions were generated by superimposing objects from MS-COCO [14] onto objects in PASCAL3D+ [15] images. Zhan *et al*. established an automatically generated subset of images from the COCO dataset with partial occlusions [17]. In multiple other works, task-specific occlusions are created in specific ways for demonstrating the efficacy of the corresponding methods. For example, in [18], a small reasonable number of clean and occluded image pairs are assumed to be available. Given a pair, a Deep-Feature-Vector (DFV) is extracted from each image. The difference between the DFV of clean and occluded image has information about the occluder. This difference vector is then added to the DFVs of all other clean images which do not have an occluded pair for the purposes of training. Training with such perturbed DFVs improved classification performance under occlusions. Other datasets for specialized tasks include the Real World Occluded Faces (ROF) dataset [19] created specifically for face recognition.

3 `LEARN` for Occlusion-Robust Classification

We first provide overall functioning of the proposed `LEARN` model, and also discuss the distinction between the current task and basic denoising or reconstruction problems. Following that, we explain the design choices and working of our loss functions towards training of the `LEARN`.

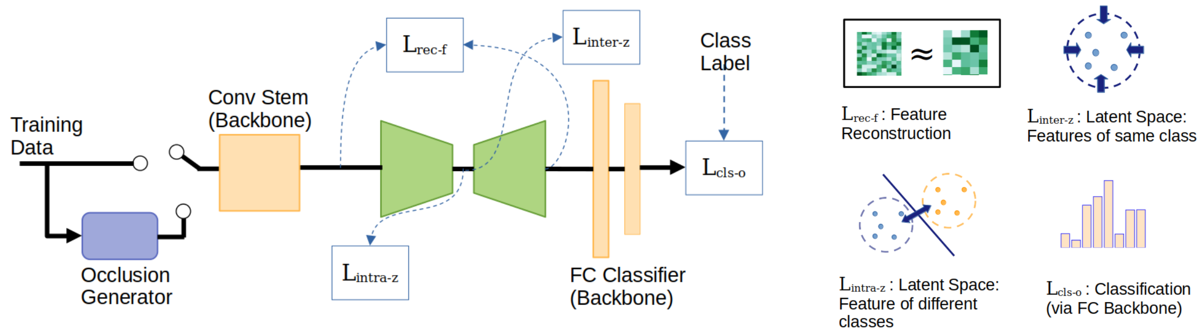


Figure 2: The schematic of the proposed LEARN: (a) shows the overall training pipeline along with loss functions, and (b) provides simple illustration of individual loss components. The green block depicts LEARN in the form of AutoEncoder.

Let $f \in \mathbb{R}^{c.h.w}$ be the output of the final convolutional layer of the classification CNN, which we also refer to as the backbone network. In a typical CNN, the corresponding output, f, is flattened and passed through one or more FC layers, where the final FC layer has dimensionality equal to the number of classes and its output indicates the likelihood of the input belonging to each class. In the following discussion, we denote $x_{m,n}$ as the m-th sample of the n-th class, and its corresponding features as $f_{m,n}$. Additionally, the occluded versions of these samples and corresponding features are denoted as $x^{\text{occ}}_{m,n}$ and $f^{\text{occ}}_{m,n}$, respectively.

To improve robustness to occlusions, working with feature representation f is often more effective than than directly working with images [5]. Likewise, we regard these feature maps, f, as the input for our AE-based LEARN where the goal is to learn the latent space for a set of inputs of one or more classes and also to reconstruct the features of occluded images for improved classification. As shown in Fig. 2, the reconstructed features are then fed back to the cascade of FC layers of the backbone network for further processing. Therefore, one may view our proposed LEARN being integrated into an existing pretrained classification backbone for the purpose of correcting or enhancing the features originating from occluded data. It is important to emphasize that enhancing features of occluded data may seem like a denoising problem, but significantly differs from processing images corrupted with noise such as AWGN or salt-and-pepper. These forms of noise affect the local neighborhood of pixels, whereas occlusion substantially degrades large spatial regions– thereby making it nearly impossible to reconstruct or learn from local-level interactions of pixels/ features. Consequently, a conventional denoising AE, trained with simple reconstruction loss, does not yield a satisfactory solution to OOD classification tasks.

For a given LEARN, we hypothesize that a joint approach to reconstruction and classification is crucial for successful classification of occluded data. It is necessary to reconstruct the features of occluded data in such a way that they can serve as improved inputs for classification– without compromising the accuracy on clean, non-occluded data. We propose a following multi-objective loss function to train the LEARN for our classification task.

Architecture of the LEARN: We use a conventional AE architecture where the encoder consists of a stack of 3 convolutional layers (conv) with the kernel size of 3. The number of output channels for each conv layer are 64, 64, and 32, respectively. Except the last conv layer (which leads to the latent space), outputs of other conv layers are passed through ReLU activation and pooling. The latent vector is passed through the decoder of the LEARN to reconstruct the features. The decoder architecture is exactly mirrored from that of the encoder by replacing conv by transposed-conv operations. The outputs of final transposed-conv layer are restricted to [-1, +1] through hard Tanh activation.

Reconstructing Occluded Features: The main objective of the LEARN is an attempt to reconstruct the features of occluded image. Initially, we create occlusions on a clean image, as a part of preprocessing and input the corresponding features ($f^{\text{occ}}_{m,n}$) for training the LEARN. The features of clean image ($f_{m,n}$) act as the reference or ground truth for reconstruction using MSE loss as shown by Equation 1.

$$\mathcal{L}_{\text{rec-f}} = \left|\left| \widetilde{f^{\text{occ}}_{m,n}} - f_{m,n} \right|\right|_2^2, \tag{1}$$

where $\tilde{f}$ indicates the reconstructed features, *i.e.*, the output of LEARN.

Constraining Intra-Class Latent Space: Due to occlusion's substantial impact on large regions of image/pixel data, majority of underlying signal (pixel data) is missing. The current trend for addressing such challenges involves the use of generative networks; however, these require massive amounts of training data and computational resources. We simplify this challenge by focusing on not attaining perfect reconstruction but instead ensuring that the reconstructed features contain sufficient class-discriminatory information. As a result, we define an auxiliary loss function to ensure that the latent space of `LEARN` does not alter while processing different samples of the same class, but enforces the latent vector of occluded samples to lie close to that of clean samples. For samples $x_{m1,n}$ and $x_{m2,n}$, we achieve these objectives by calculating the loss on pairs of clean-clean as well as clean-occluded images. The corresponding loss $\mathcal{L}_{\text{intra-z}}$ is the MSE between the corresponding latent vectors z, provided as:

$$\mathcal{L}_{\text{intra-z}} = \left|\left|z_{m1,n} - z_{m2,n}\right|\right|_2^2. \tag{2}$$

Discriminating Inter-Class Latent Space: Given the significant diversity in occluded data, the `LEARN` is vulnerable to learn non-compact, over-complete latent space for individual classes—which may degrade the classification accuracy. To address this issue, we introduce auxiliary inter-class latent loss that enforces discriminability among samples of different classes within latent space. This loss, $\mathcal{L}_{\text{inter-z}}$, formulated in a contrastive framework, is calculated on both-features of clean and occluded images of different classes as shown in Equation 3.

$$\mathcal{L}_{\text{inter-z}} = y\,(z_{m,n1} - z_{m,n2})^2 + (1-y)\,\max(0, M - (z_{m,n1} - z_{m,n2})^2), \tag{3}$$

where y indicates a class-similarity label, which is set to 1 for same classes, *i.e.* $n1 == n2$, else zero; while *M* indicates a permissible inter-class distance in the latent space.

Classification Loss: Finally, we employ the FC-layer cascade from the backbone model, along with its learnt weights, to improve the performance of the `LEARN` towards classification. The output of `LEARN` is passed to the frozen FC layer(s) to obtain a D-dimensional vector which is then used to compute the cross-entropy classification loss. The loss is backpropagated to the `LEARN` without altering the FC layers of backbone. In addition to improving the classifiability of occluded data, this loss term guides alignment of reconstructed features with the pretrained classifier of the backbone. If $\Theta(.)$ denotes the parameters of the FC-layers of the backbone, $d \in D$ refers to the class label, and *CE* is the standard cross-entropy loss, the classification loss $\mathcal{L}_{\text{cls}}$ term can be obtained as:

$$\mathcal{L}_{\text{cls-o}} = \text{CE}\left(\Theta(f_{m,n}^{\tilde{\text{occ}}}, d)\right), \tag{4}$$

We compute $\mathcal{L}_{\text{cls-o}}$ on features of both- occluded and clean data.

The overall loss function for training the `LEARN` is obtained by a weighted combination of the aforementioned loss components as shown in Equation 5. The variables $\lambda_{\_}$ refer to the relative weight of corresponding loss term.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec-f}} + \lambda_{\text{intra-z}}\mathcal{L}_{\text{intra-z}} + \lambda_{\text{inter-z}}\mathcal{L}_{\text{inter-z}} + \lambda_{\text{cls-o}}\mathcal{L}_{\text{cls-o}} \tag{5}$$

The combined loss term enables multiple strategies to synergistically address occluded features and limit the impact on latent space without compromising performance on high-quality (clean) data.

4 Experimental Results

Datasets: We evaluate the efficacy of `LEARN` on two datasets well-referenced for this task: the Pascal and MS-COCO Occluded Vehicles dataset (hereafter referred to as MS-COCO). Pascal dataset is a subset of the PASCAL3D+ dataset [15], created by extracting vehicle images and occluding them synthetically by four types of occluders: white noise, random noise, textures and lastly objects. Depending on the percentages of occluded area, the levels of occlusion can be classified into four categories: L0 (0%), L1 (20–40%), L2 (40–60%), L3 (60–80%). Whereas occlusions in the Pascal dataset are synthetic, we also consider evaluation of the proposed `LEARN` on the MS-COCO dataset [20] that contains real occlusions. Throughout this work, we have used the training split of Pascal dataset to train our models.

Backbones: For our experiments, we consider two backbone architectures, ResNet-50 and VGG16– which are pre-trained on the ImageNet-1k dataset. To obtain baseline results, we first fine-tune the backbones on the training partition of the Pascal dataset (thus, 12 classes here), following the setup described in [21].

`LEARN` **Pipeline:** The finetuned backbones (baseline models) act as the reference for our subsequent experiments. We incorporate the proposed AE-based `LEARN` after the final convolutional layer, as shown in Fig. 2 while freezing the weights of all backbone layers. To train the `LEARN` to reconstruct the feature maps of the occluded images, we generate comprehensive occlusions on-the-fly where we occlude the training samples using four occluders: white noise, random noise, texture, natural objects and the degree or level of occlusions varies from 10% to 90%. For each backbone: VGG16 and ResNet-50, we trained the model for 40 epochs (with early stopping criteria) with an initial learning rate of $1e^{-4}$ and batch size of 128.

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 1: 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. All values refer to classification accuracy in %. For the methods marked with *, we report the results from [20].

Occ. Area	L0	L1	L2	L3	Mean
baseline	99.9	94.5	83.0	55.4	79.4
RCNet [23]	99.1	96.1	86.8	59.1	85.3
RCNet++ [23]	99.4	96.8	87.2	59.2	85.7
CompNet-Res50 [24]	99.3	98.2	94.9	80.5	91.8
Proposed	**100**	**99.3**	**96.3**	**83.5**	**93.6**

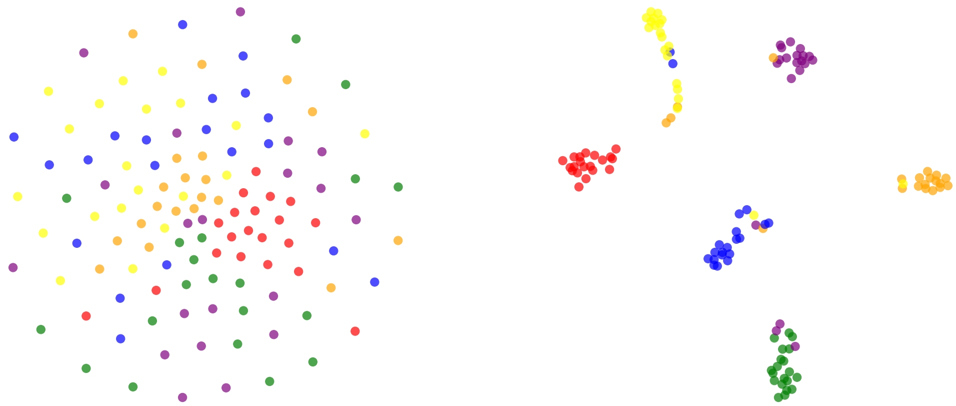
Table 2: Performance evaluation of the LEARN on the Pascal dataset, using ResNet-50 backbone, on varying levels of occlusions. All values refer to classification accuracy in %.

Results on Pascal: Table 1 shows the performance, in terms of classification accuracy, of proposed LEARN along with comparison with some of the recent state-of-the-art methods such as CompositionalNets [20], dictionary-based-compositional model [5], TDAPNet [22]. We omit the details of other methods due to brevity of space. We also include the classification accuracy on the baseline (*i.e.* the backbone finetuned on the clean Pascal dataset). For all experiments, we maintain consistency across train and test protocols used by other comparative methods—which facilitates straightforward comparison.

It can be observed that the proposed LEARN outperforms CoD, VGG+CoD and TDAPNet methods under every occlusion type and level with improvements as high as 10% for severely occluded images. Compared to the baseline, LEARN achieves an improvement of 13.8% in terms of average classification accuracy (consists of clean + OOD data). Using LEARN brings about 1–2% improvement across most occlusion types/levels compared to state-of-the-art CompNets, though, on average, the proposed method falls short of CompNets by 0.3%. One of the most challenging scenario is presented by occluding the image by another object (*e.g.* lower rows of Fig. 1). For these occlusions, LEARN outperforms every competitive method across different levels of occlusions (referred under the columns **o** in Table 1). It is important to note that the incorporation of proposed model into VGG16 backbone does not degrade the performance on clean images (level L0). While other methods resulted in nominal decrease in classification accuracy of clean images, LEARN yielded perfect classification of the same.

The results of our experiments with ResNet-50 backbone are provided in Table 2. We also enlist the results of recent state-of-the art methods which include RCNet [23] with a 3D-aware head, and CompNet [24]. In terms of average classification accuracy, LEARN improves the baseline by 14.2%. On Pascal test dataset, the LEARN outperforms all the state-of-the art methods implementing a ResNet-50 backbone by margins ranging from 1.8–8.3%. The consistent improvement in classification across all levels of occlusion demonstrates efficacy of the proposed LEARN across different architectures of classification CNNs.

To analyse the improvements brought by LEARN towards bridging the gap between intermittent features of clean and occluded images of the same class, we studied t-SNE plots of different types and levels of occlusions. Fig. 3 shows the t-SNE plots of the features (input to the classifier head) of the Pascal test dataset with the highest level of occlusion, L3 (60–80%) for the VGG16 backbone. Due to high degree of occlusions (missing/ noisy data), the intermittent features of the baseline appear highly scattered, however with the LEARN, these features are well-clustered without decreasing the inter-class margins which in-turn indicate the discriminative information of features. It should also be noted that these features (output of LEARN) are further subject to classification layers of the backbone– thus further enhancing the classification performance of the proposed model as seen in Table 1.



(a) Baseline (b) LEARN

Figure 3: t-SNE plots of the features inputted to the classifier head of the VGG16 backbone. Images with object occlusions of level L3 (60–80%) from the Pascal test dataset are considered.

Results on MS-COCO: Table 3 shows the classification results on the MS-COCO dataset for the VGG16 backbone, where the compared models were trained on the Pascal dataset. In addition to the cross-dataset testing, this dataset provides another benefit of evaluating the proposed method on realistic occlusion scenarios with varying degrees of occlusions. Similar to the previous experiments, we ensure consistent protocols to facilitate easy comparisons with the recent state-of-the-art methods of classifying occluded images. Results of classification accuracy indicate that the backbone incorporated with LEARN outperforms all compared methods for large occlusions (levels L2–L3) by at least 3.5%. With an average accuracy of 92.1%, our method provides over 13% improvements in overall classification over the baseline model. Similar to the experiments on Pascal dataset, the proposed LEARN enhances classification of clean data as well.

Occ. Area	**L0**	**L1**	**L2**	**L3**	**Mean**
baseline	98.0	83.2	76.7	57.5	78.8
CoD*	91.8	82.7	83.3	76.7	83.6
VGG+CoD*	98.0	88.7	80.7	69.9	84.3
TDAPNet*	98.0	88.5	85.0	74.0	86.4
CompNet*	98.5	**93.8**	87.6	79.5	89.9
Proposed	**99.2**	93.3	**91.1**	**84.9**	**92.1**

Table 3: Performance evaluation of LEARN on MS-COCO dataset using VGG16 backbone. All values refer to classification accuracy in %. For the methods marked with *, we report the results from [20] for consistent protocol.

Finally, we conduct the cross-dataset experiment (trained on Pascal, tested on MS-COCO) with the ResNet-50 backbone. The comparative performance, as shown in Table 4, indicates that use of LEARN brings significant improvements over baseline for each level of occlusion as well as on clean (unoccluded) images. With 89.9% accurate classifications, the LEARN results in a boost of 9.3% over the baseline, albeit falling short by 1% compared to the corresponding results obtained by the CompNet.

Occ. Area	**L0**	**L1**	**L2**	**L3**	**Mean**
baseline	97.4	82.4	77.1	65.7	80.6
CompNet [24]	98.5	**92.6**	88.9	**83.6**	**90.9**
Proposed	**98.7**	92.1	**89.5**	79.4	89.9

Table 4: Performance evaluation of LEARN on MS-COCO dataset, using ResNet-50 backbone, on varying levels of occlusions. All values refer to classification accuracy in %.

5 Conclusion

In this work, we developed a method to enhance the classification of occluded images. Our model is an AE-based network that can be integrated into common CNN-based classification models in order to improve performance on occluded and OOD data while maintaining high accuracy on clean images. The novelty of our approach lies in devising loss functions across different layers of the overall architecture, which collectively contribute to the desired improvement. In addition to conventional losses used for AE, we introduce auxiliary intra-class loss to (partially) recover latent features of the occluded data. We also employ auxiliary inter-class loss in the same latent space to ensure compact representations of the class and thereby improve inter-class margin. Our experiments on two different datasets and two backbones show that the use of `LEARN` brings as high as 25% improvements over baseline for classification of highly occluded images without compromising classification of clean (in-distribution) images.

Although our current model successfully classifies a specific set of objects (12 in this case), it may become saturated and fail to scale for a larger number of classes. We are working towards combining multiple models in order to address this issue.

References

[1] S. Gilroy, E. Jones, and M. Glavin, "Overcoming occlusion in the automotive environment–a review," *IEEE T-ITS*, vol. 22, no. 1, pp. 23–35, 2019. 1

[2] J. Kim, W. Shin, H. Park, and J. Baek, "Addressing the occlusion problem in multi-camera people tracking with human pose estimation," in *IEEE/CVF CVPR Workshops*, 2023, pp. 5463–5469. 1

[3] D. Zeng, R. Veldhuis, and L. Spreeuwers, "A survey of face recognition techniques under occlusion," *IET biometrics*, vol. 10, no. 6, pp. 581–606, 2021. 1

[4] K. Saleh, S. Szénási, and Z. Vámossy, "Occlusion handling in generic object detection: A review," in *IEEE SAMI*, 2021, pp. 000477–000484. 1

[5] A. Kortylewski, Q. Liu, H. Wang, Z. Zhang, and A. Yuille, "Combining compositional models and deep networks for robust object classification under occlusion," in *IEEE/CVF WACV*, 2020, pp. 1333–1341. 1, 3, 4, 6

[6] B. Zhao, S. Yu, W. Ma, M. Yu, S. Mei, A. Wang, J. He, A. Yuille, and A. Kortylewski, "OOD-CV: a benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images," in *ECCV*. Springer, 2022, pp. 163–180. 2

[7] B. Zhao, J. Wang, W. Ma, A. Jesslen, S. Yang, S. Yu, O. Zendel, C. Theobalt, A. Yuille, and A. Kortylewski, "OOD-CV-v2: An extended benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images," *arXiv:2304.10266*, 2023. 2

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778. 2

[9] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, vol. 25. 3

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, 09 2014. 3

[11] Z. Zhang, C. Xie, J. Wang, L. Xie, and A. Yuille, "Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion," in *IEEE/CVF CVPR*, 2018, pp. 1372–1380. 3

[12] A. Kortylewski, Q. Liu, A. Wang, Y. Sun, and A. Yuille, "Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion," *IJCV*, vol. 129, no. 3, pp. 736–760, 2021. 3

[13] A. Wang, Y. Sun, A. Kortylewski, and A. Yuille, "Robust object detection under occlusion with context-aware compositionalnets," in *IEEE/CVF CVPR*, 2020, pp. 12642–12651. 3

[14] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. 3

[15] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond PASCAL: A benchmark for 3D object detection in the wild," in *IEEE WACV*, 2014, pp. 75–82. 3, 5

[16] Y. Liu, S. Yan, L. Leal-Taixe, J. Hays, and D. Ramanan, "Soft augmentation for image classification," in *IEEE/CVF CVPR*, 2023, pp. 16241–16250. 3

[17] G. Zhan, W. Xie, and A. Zisserman, "A tri-layer plugin to improve occluded detection," *arXiv 2210.10046*, 2022. 3

[18] F. Cen, X. Zhao, W. Li, and G. Wang, "Deep feature augmentation for occluded image classification," *Pattern Recognition*, vol. 111, pp. 107737, 2021. 3

[19] M. Erakin, U. Demir, and H. Ekenel, "On recognizing occluded faces in the wild," in *BIOSIG*, 2021, pp. 1–5. 3

[20] 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. 5, 6, 7

[21] J. Wang, Z. Zhang, C. Xie, V. Premachandran, and A. Yuille, "Unsupervised learning of object semantic parts from internal states of cnns by population encoding," *arXiv:1511.06855*, 2015. 5

[22] M. Xiao, A. Kortylewski, R. Wu, S. Qiao, W. Shen, and A. L. Yuille, "TDAPNet: Prototype network with recurrent top-down attention for robust object classification under partial occlusion," *CoRR*, vol. abs/1909.03879, 2019. 6

[23] A. Jesslen, G. Zhang, A. Wang, A. Yuille, and A. Kortylewski, "Robust 3D-aware object classification via discriminative render-and-compare," *arXiv:2305.14668*, 2023. 6

[24] A. Kortylewski, Q. Liu, A. Wang, Y. Sun, and A. Yuille, "Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion," *IJCV*, vol. 129, pp. 736–760, 2021. 6, 7