3-25-2021

Orthogonal Projection Loss

Kanchana Ranasinghe
Mohamed bin Zayed University of Artificial Intelligence

Muzammal Naseer
Australian National University

Munawar Hayat
Monash University

Salman Khan
Mohamed bin Zayed University of Artificial Intelligence & Australian National University

Fahad Shahbaz Khan
Mohamed bin Zayed University of Artificial Intelligence & Linköping University

Orthogonal Projection Loss

Kanchana Ranasinghe[†], Muzammal Naseer[*], Munawar Hayat[§], Salman Khan[†*], Fahad Shahbaz Khan[†‡]
[†]Mohamed bin Zayed University of AI, UAE [*]Australian National University, Australia
[§]Monash University, Australia [‡]Linköping University, Sweden
kanchana.ranasinghe@mbzuai.ac.ae

Abstract

Deep neural networks have achieved remarkable performance on a range of classification tasks, with softmax cross-entropy (CE) loss emerging as the de-facto objective function. The CE loss encourages features of a class to have a higher projection score on the true class-vector compared to the negative classes. However, this is a relative constraint and does not explicitly force different class features to be well-separated. Motivated by the observation that ground-truth class representations in CE loss are orthogonal (one-hot encoded vectors), we develop a novel loss function termed 'Orthogonal Projection Loss' (OPL) which imposes orthogonality in the feature space. OPL augments the properties of CE loss and directly enforces inter-class separation alongside intra-class clustering in the feature space through orthogonality constraints on the mini-batch level. As compared to other alternatives of CE, OPL offers unique advantages e.g.*, no additional learnable parameters, does not require careful negative mining and is not sensitive to the batch size. Given the plug-and-play nature of OPL, we evaluate it on a diverse range of tasks including image recognition (CIFAR-100), large-scale classification (ImageNet), domain generalization (PACS) and few-shot learning (mini-ImageNet, CIFAR-FS, tiered-ImageNet and Meta-dataset) and demonstrate its effectiveness across the board. Furthermore, OPL offers better robustness against practical nuisances such as adversarial attacks and label noise. Code is available at:* https://github.com/kahnchana/opl.

1. Introduction

Recent years have witnessed great success across a range of computer vision tasks owing to progress in deep neural networks (DNNs) [24]. Effective loss functions for DNNs training have been a crucial component of these advancements [15]. In particular, the softmax cross entropy (CE) loss, commonly used for tackling classification problems, has been pivotal for stable and efficient training of DNNs.

Multiple variants of CE have been explored to enhance discriminativity and generalizability of feature representations learned during training. Contrastive [16] and triplet [45] loss functions are a common class of methods that have gained popularity on tasks requiring more discriminative features. At the same time, methods like centre loss [59] and contrastive centre loss [38] have attempted to explicitly enforce inter-class separation and intra-class clustering through Euclidean margins between class prototypes. Angular margin based losses [33, 32, 7, 55, 54] compose another class of objective functions that increase inter-class margins through altering the logits prior to the CE loss.

While these methods have proven successful at promoting better inter-class separation and intra-class compactness, they do possess certain drawbacks. Contrastive and triplet loss functions [16, 45] are dependent on carefully designed negative mining procedures, which are both time-consuming and performance-sensitive. Methods based on

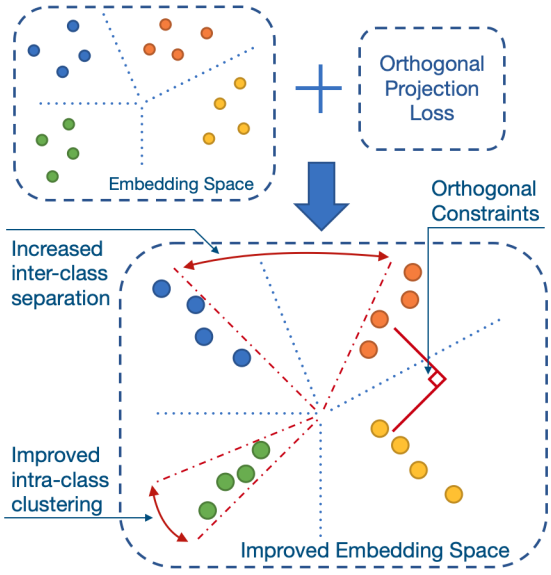


Figure 1: **Orthogonal Projection Loss:** During training of a deep neural network, within each mini-batch, OPL enforces separation between features of different class samples while clustering together features of the same class samples. OPL integrates well with softmax CE loss as it simply complements its intrinsic angular property, leading to consistent performance improvements on various classification tasks with a variety of DNN backbones.

centre loss [59, 38], that work together with CE loss, promote margins in Euclidean space which is counter-intuitive to the intrinsic angular separation enforced through CE loss [32]. Further, these methods introduce additional learnable parameters in the form of new class centres. Angular margin based loss functions [32, 33] which are highly successful for face recognition tasks, make strong assumptions for face embeddings to lie on the hypersphere manifold, which does not hold universally for all computer vision tasks [47]. Some loss designs are also specific to certain architecture classes *e.g.*, [47] can only work with DNNs which output Class Activation Maps [66].

In this work, we explore a novel direction of simultaneously enforcing inter-class separation and intra-class clustering through orthogonality constraints on feature representations learned in the penultimate layer (Fig. 1). We propose Orthogonal Projection Loss (OPL), which can be applied on the feature space of any DNN as a plug-and-play module. We are motivated by how image classification inherently assumes independent output classes and how orthogonality constraints in feature space go hand in hand with the one-hot encoded (orthogonal) label space used with CE. Furthermore, orthogonality constraints provide a definitive geometric structure in comparison to arbitrarily increasing margins which are prone to change depending on the selected batch, thus reducing sensitivity to batch composition. Finally, simply maximizing the margins can cause negative correlation between classes and thereby unnecessarily focus on well-separated classes while we tend to ensure independence between different class features to successfully disentangle the class-specific characteristics.

Compared with contrastive loss functions [16, 45], OPL operates directly on mini-batches, eliminating the requirement of complex negative sample mining procedures. By enforcing orthogonality through computing dot-products between feature vectors, OPL provides a natural augmentation to the intrinsic angular property of CE, as opposed to methods [59, 38, 17] that enforce an Euclidean margin in feature space. Furthermore, OPL introduces no additional learnable parameters unlike [59, 53, 38], operates independent of model architecture unlike [47], and in contrast to losses operating on the hypersphere manifold [32, 33, 7, 55], performs well on a wide range of tasks. Our main contributions are:

- We propose a novel loss, OPL, that directly enforces inter-class separation and intra-class clustering via orthogonality constraints with no learnable parameters.
- Our orthogonality constraints are efficiently formulated compared to existing methods [27, 48], allowing mini-batch processing without the need to explicitly obtain singular values. This leads to a simple vectorized implementation of OPL directly integrating with CE.
- We extensively evaluate on a diverse range of image classification tasks highlighting the discriminative ability of OPL. Further, our results on few-shot learning (FSL) and domain generalization (DG) datasets establish the transferability and generalizability of features learned with OPL. Finally, we establish the improved robustness of learned features to adversarial attacks and label noise.

2. Related Work

Loss Functions. Loss functions play a central role in all deep learning based computer vision tasks. Recent works include reformulating the generic softmax calculation to limit the variance and range of logits during training [64], constraining the feature space to follow specific distributions [53], and heuristically altering margins targeting specific tasks like few-shot learning [31, 28], class imbalance [20, 21, 17, 30] and zero-shot learning [39]. OPL optimizes a different objective of inter-class separation and intra-class clustering through orthogonalization in the feature space.

Generalizable Representations. Recent works explore the transferability of features learned via supervised training [62], *e.g.* FSL [51, 4, 14] and DG [19, 9] tasks. Tian *et al.* [51] establish a strong FSL baseline using only standard (non-episodic) supervised pre-training. Adaptation of supervised pre-trained models to the episodic evaluation setting of FSL tasks is explored in [61, 4]. Goldblum *et al.* [14] show the importance of margin-based regularization methods for FSL. Our work differs by building on orthogonality constraints to learn more transferable features, and is more compatible with CE as opposed to [14]. Multiple DG methods also explore constraints on feature spaces [19, 9] to boost cross-domain performance. In particular, [9] explores inter-class separation and intra-class clustering through contrastive and triplet loss functions. OPL improves on these while eliminating the need for compute expensive and complex sample mining procedures.

Orthogonality. Orthogonality of kernels in DNNs is well explored with an aim to diversify the learned weight vectors [34, 56]. The idea of orthogonality is also used for disentangled representations such as in [57] and to stabilize network training since orthogonalization ensures energy preservation [8, 43]. Orthogonal weight initializations have also shown their promise towards improving learning behaviours [37, 60]. However, all of these works operate in the parameter space. Remarkably, the previous formulations to achieve orthogonality in the feature space generally depend on computing singular value decomposition [27, 48], which can be numerically unstable, difficult to estimate for rectangular matrices, and undergoes an iterative process [2]. In contrast, our orthogonal constraints are enforced in a novel manner, realized via decomposition on the sample-to-sample relationships within a mini-batch, while simultaneously avoiding tedious pair/triplet computations.

3. Proposed Method

Maximizing inter-class separation while enhancing intra-class compactness is highly desirable for classification. While the commonly used cross entropy (CE) loss encourages logits of the same class to be closer together, it does not enforce any margin amongst different classes. There have been multiple efforts to integrate max-margin learning with CE. For example, Large-margin softmax [33] enforces inter-class separability directly on the dot-product similarity while SphereFace [32] and ArcFace [7] enforce multiplicative and additive angular margins on the hypersphere manifold, respectively. Directly enforcing max-margin constraints to enhance discriminability in the angular domain is ill-posed and requires approximations [33]. Some works turn to the Euclidean space to enhance feature space discrimination. For example, centre loss [59] clusters penultimate layer features using Euclidean distance. Similarly, Affinity Loss [17] forms uniformly shaped equidistant class-wise clusters based upon Gaussian distances in the Euclid space. Margin-maximizing objective functions in the Euclid space are not ideally suited to work alongside CE loss, since CE seeks to separate output logits in the angular domain. By enforcing orthogonality constraints, our proposed OPL loss maximally separates intermediate features in the angular domain, thus complementing cross-entropy loss which enhances angular discriminability in the output space. In the following discussion, we revisit CE loss in the context of max-margin learning, and argue why OPL loss is ideally suited to supplement CE.

3.1. Revisiting Softmax Cross Entropy Loss

Consider a deep neural network $\mathcal{H}$, which can be decomposed into $\mathcal{H} = \mathcal{H}_\phi \cdot \mathcal{H}_\theta$, where $\mathcal{H}_\phi$ is the feature extraction module and $\mathcal{H}_\theta$ is the classification module. Given an input-output pair $\{\mathbf{x}, y\}$, let $\mathbf{f} = \mathcal{H}_\phi(\mathbf{x}), \mathbf{f} \in \mathbb{R}^d$ be the intermediate features and $\hat{y} = \mathcal{H}_\theta(\mathbf{f}), \hat{y} \in \mathbb{R}^k$ be the output predictions. For brevity, let us define the classification module as a linear layer $\mathcal{H}_\theta = \mathbf{W} = [\mathbf{w}_1, \cdots \mathbf{w}_c]$ with no unit-biases, where $\mathbf{w}_i \in \mathbb{R}^d, i = 1 \cdots c$ are class-wise learnable projection vectors for c classes. The traditional CE loss can then be defined in terms of discrepancy between the predicted $\hat{y}$ and ground-truth label y, by projecting the features $\mathbf{f} \in \mathbb{R}^d$ onto the weight matrix $\mathbf{W} \in \mathbb{R}^{d \times c}$.

$$\mathcal{L}_{\text{CE}}(\hat{y}, y) = -\log \frac{\exp\left(\mathbf{f}^T \mathbf{w}_y\right)}{\sum_j \exp\left(\mathbf{f}^T \mathbf{w}_j\right)} \tag{1}$$

$$\propto \sum_{j \neq y} \exp\left(\mathbf{f}^T \mathbf{w}_j - \mathbf{f}^T \mathbf{w}_y\right)$$

$$\propto \sum_{j \neq y} \exp\left(\|\mathbf{f}\|_2 \|\mathbf{w}_j\|_2 \cos(\theta_j) - \|\mathbf{f}\|_2 \|\mathbf{w}_y\|_2 \cos(\theta_y)\right)$$

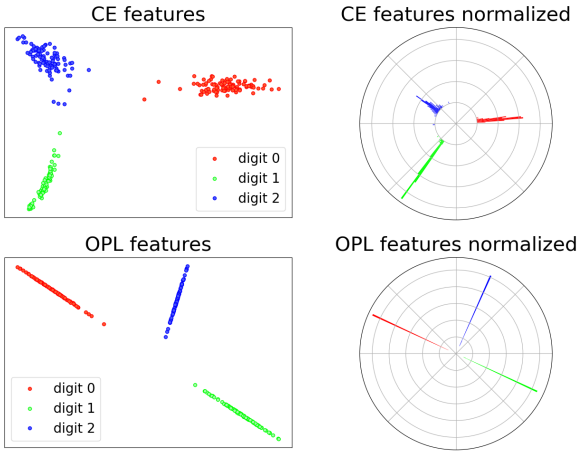


Figure 2: **Feature space visualization for CE vs OPL:** Inter-class orthogonality enforced by OPL can be observed in this MNIST 2-D feature visualization. We plot only three classes to better illustrate in 2D the inter-class orthogonality achieved. Normalization refers to projection of vectors to a unit-hypersphere in feature-space, and the normalized plot contains a histogram for each angle.

Since CE does not explicitly enforce any margin between each class pair, previous works have shown that the learned class regions for some class samples tend to be bigger compared to others [32, 7, 17]. To counter this effect and ensure all classes are equally separated, efforts have been made to introduce a margin m between different classes by modifying the term $\cos(\theta_y)$ as $\cos(m + \theta_y)$ for additive angular margin [7], $\cos(m\theta_y)$ for multiplicative angular margin [32] and $m + \cos(\theta_y)$ as additive cosine margin [55]. The gradient propagation for such margin-based softmax loss formulations is hard, and previous works rely on approximations. Instead of introducing any margins to ensure uniform seperation between different classes, our proposed loss function simply enforces all classes to be orthogonal to each other with simultaneous clustering of within-class samples, using an efficient vectorized implementation with straightforward gradient computation and propagation.

By considering the $\mathbf{w}_y \in \mathbf{W}$ vectors as individual class prototypes, the CE loss can be viewed as aligning the feature vectors $\mathbf{f}$ along its relevant class prototype. The cosine similarity in the form of the dot product ($\mathbf{f}^T \mathbf{w}_y$) gives CE an intrinsic angular property, which is observed in Fig. 2 where features naturally separate in the polar coordinates with CE only. Moreover, during the standard SGD based optimization, the CE loss is applied on mini-batches. We note that there is no explicit enforcement of feature separation or clustering across multiple samples within the mini-batch. Given the opportunity to enforce such constraints since supervised training is commonly conducted adopting random mini-batch based iterations, we explore the possibility of within mini-batch constraints aimed at augmenting the intrinsic discriminative characteristics of CE loss.

3.2. Orthogonal Projection Loss

The CE loss with one-hot-encoded ground-truth vectors seeks to implicitly achieve orthogonality between different classes in the output space. Our proposed OPL loss, ameliorates CE loss, by enforcing class-wise orthogonality in the intermediate feature space. Given an input-output pair $\{\mathbf{x}_i, y_i\}$ in the dataset $\mathcal{D}$, let $\mathbf{f}_i = \mathcal{H}_\phi(\mathbf{x}_i)$ be the features output by an intermediate layer of the network. Our objective is to enforce constraints to cluster the features $\mathbf{f}_i \ \forall \mathbf{x}_i \in \mathcal{D}$ such that the features for different classes are orthogonal to each other and the features for the same class are similar. To this end, we define a unified loss function that simultaneously ensures intra-class clustering and inter-class orthogonality within a mini-batch as follows:

$$s = \sum_{\substack{i,j \in B \\ y_i = y_j}} \langle \mathbf{f}_i, \mathbf{f}_j \rangle \tag{2}$$

$$d = \sum_{\substack{i,k \in B \\ y_i \neq y_k}} \langle \mathbf{f}_i, \mathbf{f}_k \rangle \tag{3}$$

$$\mathcal{L}_{\texttt{OPL}} = (1 - s) + |d| \tag{4}$$

where $\langle \cdot, \cdot \rangle$ is the cosine similarity operator applied on two vectors, $|\cdot|$ is the absolute value operator, and B denotes mini-batch size. Note that the cosine similarity operator used in Eq. 2 and 3 involves normalization of features (projection to a unit hyper-sphere) as follows:

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \cdot \|\mathbf{x}_j\|_2} \tag{5}$$

where $||\cdot||_2$ refers to the ℓ_2 norm operator. This normalization is key to aligning the outcome of OPL with the intrinsic angular property of CE loss.

In Eq. 4, our objective is to push s towards 1 and d towards 0. Since $1 - s > 0$ already, we take the absolute value of d given $d \in (-1, 1)$. This in turn restricts the overall loss such that $\mathcal{L}_{\texttt{OPL}} \in (0, 3)$. When minimizing this overall loss, the first term $(1 - s)$ will ensure clustering of same class samples, while the second term $|d|$ will ensure the orthogonality of different class samples. The loss can be implemented efficiently in a vectorized manner on the mini-batch level, avoiding any loops (see Algorithm 1).

We further note that the ratio of contribution to the overall loss of each individual term in Eq. 4 can be controlled to re-prioritize between the two objectives of inter-class separation and intra-class compactness. While the unweighted combination of s and d alone performs well, specific use-cases could benefit from weighted combinations. We reformulate Eq. 4 as follows:

$$\mathcal{L}_{\texttt{OPL}} = (1 - s) + \gamma * |d| \tag{6}$$

where γ is the hyper-parameter controlling the weight for the two different constraints.

Algorithm 1 Pytorch style pseudocode for OPL

```
def forward(features, labels):
    """
    features:   features shaped (B, D)
    labels:     targets shaped (B, 1)
    """
    features = F.normalize(features, p=2, dim=1)

    # masks for same and diff class features
    mask = torch.eq(labels, labels.t())
    eye = torch.eye(mask.shape[0])
    mask_pos = mask.masked_fill(eye, 0)
    mask_neg = 1 - mask

    # s & d calculation
    dot_prod = torch.matmul(features, features.t())
    pos_total = (mask_pos * dot_prod).sum()
    neg_total = torch.abs(mask_neg * dot_prod).sum()
    pos_mean =  pos_total / (mask_pos.sum() + 1e-6)
    neg_mean = neg_total / (mask_neg.sum() + 1e-6)

    # total loss
    loss = (1.0 - pos_mean) + neg_mean

    return loss
```

Since OPL acts only on intermediate features, we apply cross entropy loss over the outputs of the final classifier $\mathcal{H}_\theta$. The overall loss used is a weighted combination of CE and OPL. We note that our proposed loss can also be used together with other common image classification losses, such as Guided Cross Entropy, Label Smoothing or even task specific loss functions in different computer vision tasks. The overall loss $\mathcal{L}$ can be defined as:

$$\mathcal{L} = \mathcal{L}_{\texttt{CE}} + \lambda \cdot \mathcal{L}_{\texttt{OPL}} \tag{7}$$

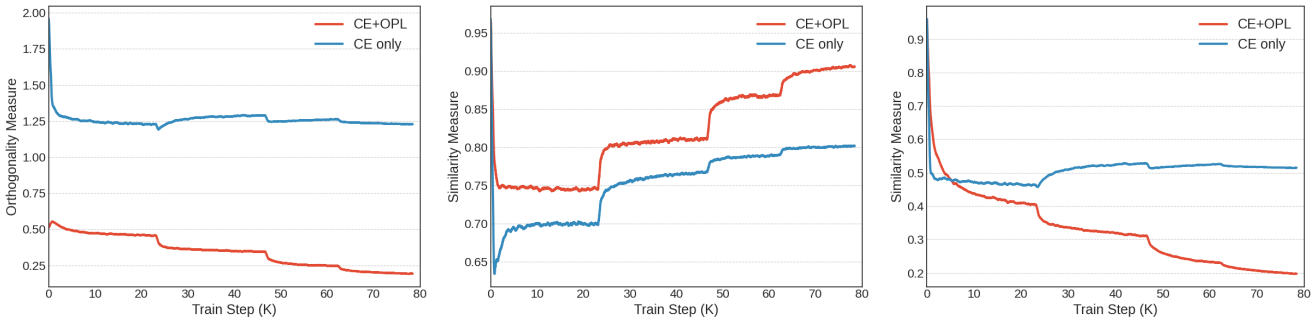
where λ is a hyper-parameter controlling the OPL weight.

3.3. Interpretation and Analysis

Overall Objective: Consider $\mathbf{F}_c$ is a set of mini-batch samples comprising of normalized features from the same class c in a given dataset $\mathcal{D}$. The overall OPL constraints can be viewed as a minimization of the following objective to update the network $\mathcal{H}_\phi$ over the random variables $\mathbf{F}_c$:

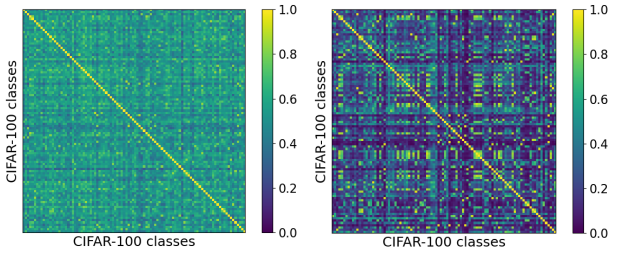
$$\min_{\phi} \sum_{i=1}^{c} \sum_{j=1}^{c} \left| \mathop{\mathbb{E}}_{(\mathbf{F}_i, \mathbf{F}_j) \sim \mathcal{D}} \left[\mathbf{F}_i \mathbf{F}_j^T \right] - [\![i = j]\!] \right|, \tag{8}$$

where $|\cdot|$ is the absolute value operator and $[\![\cdot]\!]$ is the Iverson bracket operator. We refer to the term defined in Eq. 8 as the expected inter-class orthogonality. The behaviour of OPL in terms of minimizing this expectation is visualized in Fig. 4 where average per-class feature vectors over the CIFAR-100 dataset are calculated for ResNet-56 models trained under 'CE-only' and 'CE+OPL' settings. Clear improvements over the CE baseline in terms of minimizing the expected inter-class orthogonality can be observed. Moreover, the stochastic mini-batch based application of OPL prevents naively pushing all non-diagonal values to zero as observed. This translates to allowing necessary inter-class



(a) Feature orthogonality (↓) (b) Similarity of same class features (↑) (c) Similarity of different class features (↓)

Figure 3: **Feature Analysis:** We compare feature orthogonality as measured by OPL and feature similarity as measured by cosine similarity and plot their convergence during training. Feature similarity is initially high because all features are random immediately after initialization. OPL simultaneously enforces higher inter-class similarity and intra-class dissimilarity in comparison with the CE baseline.



(a) 'CE-only' trained features (b) 'CE+OPL' trained features

Figure 4: **Orthogonality Visualization:** We present matrices illustrating the orthogonality of average per class features computed over the CIFAR-100 test set. See Appendix B.3 for more analysis.

relationships not encoded in one-hot labels of a dataset to be captured within the learned features.

Decomposing OPL: Taking a step further, we decompose OPL into its sub-components, s and d, as defined in Eq. 4. While s computes the pair-wise cosine similarity between all same-class features within the mini-batch, d calculates the similarity between different class features. These measures can directly be adopted to quantify the inter-class separation and intra-class compactness in any given features space. Moreover, the unweighted OPL formulation in Eq. 4 can be considered a measure of the overall feature orthogonality within any given embedding space. It will be interesting to compare the contribution of OPL towards inter-class separation and intra-class clustering of a feature space in contrast to the generic CE based training scenario. We present this comparison by training ResNet-56 on CIFAR-100 dataset in Fig. 3. This separation of features achieved through OPL translates to performance improvements not only within the standard classification setting, but also in tasks requiring transferable or generalizable features. Goldblum *et al*. [14] explore the significance of inter-class separation and intra-class clustering for better performance when transferring a feature embedding to a few-shot learning task. Similar notions regarding discriminative features are explored in [9] for domain generalization. We explore the effects of OPL in few-shot learning settings, and visualize the novel class embeddings learned with OPL in Appendix B.1 using LDA [36] to preserve the inter-class to intra-class variance ratio as suggested in [14].

Why orthogonality constraints?: One may wonder what benefit orthogonality in the feature space can provide in comparison to simply maximizing margin between classes. Our reasoning is twofold: reducing sensitivity to batch composition and avoiding negative correlation constraints. Within the random mini-batch based training setting, the orthogonality objective provides a definitive geometric structure irrespective of the batch composition, while the optimal max-margin separation is dependent on the batch composition. Furthermore, in the common case where the output space feature dimension $d > c$ (c number of classes), maximizing angular margin between normalized features on a unit hyper-sphere will lead to *negative* correlation among the class prototypes (considering a maximal and equi-angular separation). We argue that this is an undesired constraint since the categorical classification task itself assumes non-existence of ordinal relationships between classes (*e.g.* use of orthogonal one-hot encoded labels). Moreover, extending the constraints to additionally cause negative correlation between classes unnecessarily focuses on already well-separated classes during training whereas our constraint which tends to ensure independence provides a more balanced objective to disentangle the class-specific characteristics of even more fine-grained classes.

4. Experiments

We extensively evaluate our proposed loss function on multiple tasks including *image classification* (Tables 1 & 2), *robustness against label noise* (Table 3), *robustness against adversarial attacks* (Table 4) and *generalization to domain shifts* (Table 5). We further observe the enhanced *transferability* of orthogonal features *e.g.* in the case of few shot learning (Tables 6 & 7). Our approach shows consistent

improvements and highlights the advantages of orthogonal features on this diverse set of tasks and datasets with various deep network backbones. Additionally, we demonstrate the plug-and-play nature of OPL by showing benefits of its use over CE, Truncated Loss (for noisy labels) [65], RSC [19] and various adversarial learning baselines.

4.1. Image Classification

We evaluate the effectiveness of orthogonal features in the penultimate layer on image classification using our proposed training objective (Eq. 7). Competitive results are achieved showing consistent improvements (Tables 1 & 2) on two datasets: CIFAR-100 [22], and ImageNet [23].

CIFAR-100 consists of 60,000 natural images spread across 100 classes, with 600 images per class. We apply OPL over a cross-entropy baseline for supervised classification on CIFAR-100 (following the experiment setup in [47]), and compare our results against other loss functions which impose margin constraints [32, 33, 55, 7], introduce regularization [47, 27, 30, 6], or promote clustering [59, 64] to enhance separation among classes in Table 1. Despite its simplicity, our method performs well against state-of-the-art loss functions. Note that HNC [47] is dependant on class activation maps, RBF [64] and LGM [53] involve learnable parameters, and CB Focal Loss [6] specifically solves class-imbalance. In contrast, OPL has a simple formulation easily integratable to any network architecture, involves no learnable parameters, and targets general classification. Additionally, we note how OPL has higher performance gains with respect to top-1 accuracy (in comparison to top-5 accuracy) which is the more challenging metric. We attribute this to the fact that increased separation through OPL mostly helps in classifying difficult samples. Further, we note top-5 is not a preferred measure for CIFAR-100 since most classes are different in nature as opposed to e.g., ImageNet with several closely related classes (where our gain is much pronounced for top-5 accuracy, as discussed next).

ImageNet is a standard large-scale dataset used in visual recognition tasks, containing roughly 1.2 million training images and 50,000 validation images. We experiment with OPL by integrating it to common backbone architectures used in image classification tasks: ResNet18 and ResNet50. We train [1] the models for 90 epochs using SGD with momentum (initial learning rate 0.1 decayed by 10 every 30 epochs). Results for these experiments are presented in Table 2 and Fig. 5. We note that simply enforcing our orthogonality constraints increases the top-1 (%) accuracy of ResNet50 from 76.15% to 76.98% without any additional bells and whistles. Moreover, given the large number of fine-grained classes among the 1000 categories of ImageNet (*e.g.*, multiple dog species) which can be viewed as difficult cases, the better discriminative features learned by OPL obtains notable improvements in top-5 accuracy as well.

[1] https://github.com/pytorch/examples/tree/master/imagenet

Loss	Resnet-56		ResNet-110	
	Top-1	Top-5	Top-1	Top-5
Center Loss (ECCV'16) [59]	72.72%	93.06%	74.27%	93.20%
Focal Loss (ICCV'17) [30]	73.09%	93.07%	74.34%	93.34%
A-Softmax (CVPR'17) [32]	72.20%	91.28%	72.72%	90.41%
LMC Loss (CVPR'17) [55]	71.52%	91.64%	73.15%	91.88%
OLE Loss (CVPR'18) [27]	71.95%	92.52%	72.70%	92.63%
LGM Loss (CVPR'18) [53]	73.08%	93.10%	74.34%	93.06%
Anchor Loss (ICCV'19) [44]	-	-	74.38%	92.45%
AAM Loss (CVPR'19) [7]	71.41%	91.66%	73.72%	91.86%
CB Focal Loss (CVPR'19) [6]	73.09%	93.07%	74.34%	93.34%
HNC (ECCV'20) [47]	73.47%	**93.29%**	74.76%	**93.65%**
RBF (ECCV'20) [64]	73.36%	92.94%	-	-
CE (Baseline)	72.40%	92.68%	73.79%	93.11%
CE+OPL (Ours)	**73.52%**	93.07%	**74.85%**	93.32%

Table 1: **CIFAR-100**: These results indicate that a simple combination of cross-entropy along with our proposed orthogonal constraint gives improvements over the baseline loss function.

Method	ResNet-18		ResNet-50	
	top-1	top-5	top-1	top-5
CE (Baseline)	69.91%	89.08%	76.15%	92.87%
CE + OPL (ours)	**70.27%**	**89.60%**	**76.98%**	**93.30%**

Table 2: **Results on ImageNet:** OPL gives an improvement over a cross-entropy (CE) baseline for common backbone architectures.

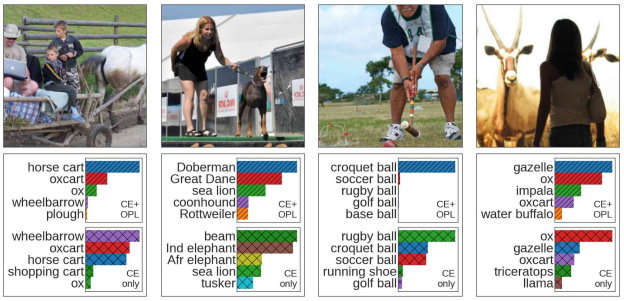


Figure 5: **Qualitative Results:** We present the top-5 predictions for OPL and CE in images where training with OPL has fixed the incorrect prediction by 'CE only' model. See Appendix B.2.

4.2. Robustness against Label Noise

Given the rich representation capacity of deep neural networks, especially considering how most can even fit random labels or noise perfectly [63], errors in the sample labels pose a significant challenge for training. In most practical applications, label noise is almost impossible to avoid, in particular, when it comes to large-scale datasets requiring millions of human annotations. Multiple works [13, 65] explore modifications to common objective functions aimed at building robustness to label noise. Despite the explicit inter-

Dataset	Method	Uniform	Class Dependent
CIFAR10	TL (NeurIPS'18) [65]	87.62%	82.28%
	TL[65] + OPL	**88.45%**	**87.02%**
CIFAR100	TL (NeurIPS'18) [65]	62.64%	47.66%
	TL[65] + OPL	**65.62%**	**53.94%**

Table 3: **Results on CIFAR-100 for Noisy Labels:** We explore the effect of noisy labels when training with OPL for image classification tasks. We use the method in [65] as a baseline comparison with 0.4 noise level and ResNet18 backbone.

class separation constraints on the feature space enforced by OPL, we argue that the random mini-batch based optimization exploited by OPL negates the effects of noisy labels. This hypothesis is supported by our experiments presented in Table 3 which show additional robustness of OPL against label noise. We simply integrate OPL over the approach followed in [65], without any task-specific modifications.

4.3. Robustness against Adversarial Attacks

Adversarial attacks modify a given benign sample by adding adversarial noise such that the deep neural network is deceived [50]. Adversarial examples are out-of-distribution samples and remain a challenging problem to solve. Adversarial training [35] emerges as an effective defense where adversarial examples are generated and added into the training set. We enforced orthogonality on such adversarial examples in the feature space while optimizing the model weights and show our benefit on different adversarial training mechanisms [35, 18, 58]. Important to note that all the considered adversarial training schemes [35, 18, 58] are different in nature *e.g.* training in Madry *et al.* [35] is based on cross-entropy only, Hendrycks *et al.* [18] propose to exploit pre-training, while Wang *et al.* [58] introduce a surrogate loss along with cross-entropy. Our orthogonality constraint help maximizing adversarial robustness in all cases showing the generic and plug-and-play nature of our proposed loss. In order to have reliable evaluation, we report robustness gains against Auto-Attack (AA) [5] in Table 4. On CIFAR10, our method increased robustness of [35] by 5.11%, [18] by 0.81% and [58] by 2.2%.

4.4. Domain Generalization (DG)

The DG problem aims to train a model using multi-domain source data such that it can directly generalize to new domains without the need of retraining. We argue that the feature space constraints of OPL tend to capture more general semantic features in images which generalize better across domains. This is verified by the performance improvements for DG that we obtain by integrating OPL with the state-of-the-art approach in [19] and evaluating on the popular PACS dataset [29]. The results presented in Table 5 indicate that integrating OPL with [19] sets new state-of-the-art across all four domains as compared to [19].

Dataset	Method	Clean	Advers.
CIFAR10	Madry *et al.*(ICLR'18) [35]	87.14	44.04
	Madry *et al.* [35] + OPL	**87.76**	**49.15**
	Hendrycks *et al.*(PMLR'19) [18]	87.11	54.92
	Hendrycks *et al.* [18] + OPL	**87.51**	**55.73**
	MART [58] (ICLR'20)	**84.49**	54.10
	MART[58] + OPL	84.41	**56.23**
CIFAR100	Madry *et al.*(ICLR'18) [35]	60.20	20.60
	Madry *et al.* [35] + OPL	**61.13**	**23.01**
	Hendrycks *et al.*(PMLR'19) [18]	59.23	28.42
	Hendrycks *et al.* [18] + OPL	**61.00**	**30.05**
	MART (ICLR'20) [58]	**58.90**	23.40
	MART[58] + OPL	58.01	**25.74**

Table 4: **OPL performance on Adversarial Robustness**: We show the impact of enforcing orthogonality on the robust features. We adversarially train baseline methods [35, 18, 58] by adding OPL constraint during training. Robust features obtained with OPL leads to better accuracy and show clear improvements over the baseline. Top-1 accuracy is reported against Auto-Attack [5] in whitebox setting (attacker has full knowledge of the model architecture and pretrained weights).

Method	Art	Cartoon	Sketch	Photo	Avg
JiGen(CVPR'19) [3]	86.20	78.70	70.63	97.66	83.29
MASF(NeurIPS'19) [10]	82.89	80.49	72.29	95.01	82.67
MetaReg(NeurIPS'18) [1]	87.20	79.20	70.30	97.60	83.60
RSC(ECCV'20) [19]	87.89	82.16	83.35	96.47*	87.47
RSC + OPL	**88.28**	**84.64**	**84.17**	**96.83**	**88.48**

Table 5: **Results on PACS dataset:** We integrate OPL with [19], gaining improvements for domain generalization tasks (*best replicated value).

4.5. Few Shot Learning (FSL)

In this section, we explore the transferability of features learned with our loss function in relation to FSL tasks. We evaluate OPL on three benchmark few-shot classification datasets: miniImageNet, tieredImageNet, and CIFAR-FS. We run additional experiments on Meta-Dataset [52] which is a large-scale benchmark for evaluating FSL methods in more diverse and challenging settings. Similar to [42], we expand Meta-Dataset by adding three additional datasets, MNIST, CIFAR10, and CIFAR100. In light of work that shows promise of learning strong features for FSL [51], we experiment with OPL using it as an auxiliary loss on the feature space during the supervised training. Quantitative results highlighting the performance improvements are presented in Table 6. Our results on Meta-Dataset are obtained using the *train on all* setting presented in [52]. We integrate OPL over the method presented in [11], train on the first 8 datasets of Meta-Dataset, and evaluate on the rest (including the three additional datasets from [42]). Results are presented in Table 7. See Appendix A.3 for robustness of OPL features against *sample noise* in FSL tasks.

Method	New Loss	Cifar:1shot	Cifar:5shot	Mini:1shot	Mini:5shot	Tier:1shot	Tier:5shot
MAML$_{(PMLR'17)}$ [12]	-	58.90±1.9	71.50±1.0	48.70±1.84	63.11±0.92	51.67±1.81	70.30±1.75
PN $_{(NIPS'17)}$ [46]	-	55.50±0.7	72.00±0.6	49.42±0.78	68.20±0.66	53.31±0.89	72.69±0.74
RN$_{(CVPR'18)}$ [49]	-	55.00±1.0	69.30±0.8	50.44±0.82	65.32±0.70	54.48±0.93	71.32±0.78
Shot-Free$_{(ICCV'19)}$ [41]	-	69.20±N/A	84.70±N/A	59.04±N/A	77.64±N/A	63.52±N/A	82.59±N/A
MetaOptNet$_{(CVPR'19)}$ [26]	-	72.60±0.7	84.30±0.5	62.64±0.61	78.63±0.46	65.99±0.72	81.56±0.53
RFS$_{(ECCV'20)}$[51]	-	71.45±0.8	85.95±0.5	62.02±0.60	79.64±0.44	69.74±0.72	84.41±0.55
RFS + OPL (Ours)	✓	**73.02±0.4**	**86.12±0.2**	**63.10±0.36**	**79.87±0.26**	**70.20±0.41**	**85.01±0.27**
NAML$_{(CVPR'20)}$ [28]	✓	-	-	65.42±0.25	75.48±0.34	-	-
Neg-Cosine$_{(ECCV'20)}$ [31]	✓	-	-	63.85±0.81	81.57±0.56	-	-
SKD$_{(Arxiv'20)}$ [40]	✓	74.50±0.9	88.00±0.6	65.93±0.81	83.15±0.54	71.69±0.91	86.66±0.60
SKD + OPL (Ours)	✓	**74.94±0.4**	**88.06±0.3**	**66.90±0.37**	**83.23±0.25**	**72.10±0.41**	**86.70±0.27**

Table 6: **Few-Shot Learning Improvements:** We obtain performance improvements using OPL over the RFS [51] baseline and SKD baseline [40] containing ResNet-12 backbones. Our loss is simply plugged in to their supervised feature learning phase. Results reported for our experiment are averaged over 3000 episodic runs. Note that [40, 28, 31] are recent loss functions specific to FSL.

Dataset	CNAPs [42] (NeurIPS'19)	SUR [11] (ECCV'20)	SUR + OPL (Ours)
Imagenet	52.3±1.0	56.4±1.2	**56.5±1.1**
Omniglot	88.4±0.7	88.5±0.8	**89.8±0.7**
Aircraft	**80.5±0.6**	79.5±0.8	79.6±0.7
Birds	72.2±0.9	76.4±0.9	**76.9±0.7**
Textures	58.3±0.7	**73.1±0.7**	72.7±0.7
Quick Draw	72.5±0.8	75.7±0.7	**75.7±0.7**
Fungi	47.4±1.0	48.2±0.9	**50.1±1.0**
VGG Flower	86.0±0.5	90.6±0.5	**90.9±0.5**
MSCOCO	42.6±1.1	**52.1±1.0**	52.0±1.0
MNIST	92.7±0.4	93.2±0.4	**94.3±0.4**
CIFAR10	61.5±0.7	66.4±0.8	**66.6±0.7**
CIFAR100	50.1±1.0	57.1±1.0	**57.6±1.0**
Average	67.0	71.4	**71.9**

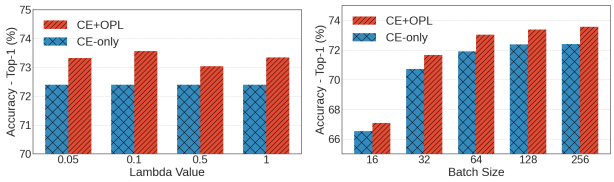
Table 7: **Results on Meta-Dataset:** OPL is integrated with the SUR-PNF method in [11] for the Meta-Dataset train on all setting. *Traffic Signs* dataset has been omitted in comparisons due to an error in Meta-Dataset possibly affecting prior work.

hyper-parameter	γ=2	γ=1	γ=0.5
λ = 0.05	70.48	70.66	72.02
λ = 0.1	70.12	70.94	71.30
λ = 0.5	70.26	71.18	70.66
λ = 1	69.78	70.48	**72.20**
λ = 2	67.64	69.58	70.52

Table 8: **Hyper-parameter search:** We report the top-1 accuracy values on a held-out validation set on CIFAR-100 using ResNet-56 backbone after training using OPL with various pairs of λ and γ hyper-parameters.

4.6. Ablative Study

OPL in its full form (Eq. 6) contains two hyper-parameters, λ and γ. We conduct a hyper-parameter search over a held-out validation set of CIFAR-100 (see Table 8).



(a) Sensitivity to λ (b) Variation with Batch Size

Figure 6: **Ablative Study:** OPL achieves consistent performance improvements against a CE-only baseline when evaluated on CIFAR-100 dataset with a ResNet-56 backbone.

The optimum values selected from these experiments are kept consistent across all other tasks and used when reporting test performance. Furthermore, we evaluate the performance of OPL on the test split of CIFAR-100 for varying λ values keeping γ fixed, to illustrate the minimal sensitivity of our method to different λ values in Fig. 6a.

Next, we consider how OPL operates on random mini-batches, and evaluate its performance against varying batch sizes (CIFAR-100 dataset). These results presented in Fig. 6b exhibit how OPL consistently

5. Conclusion

We present a simple yet effective loss function to enforce orthogonality on the output feature space and establish its performance improvements for a wide range of classification tasks. Our loss function operates in conjunction with the softmax CE loss, and can be easily integrated with any DNN. We also explore a variety of characteristics of the features learned with OPL illustrating its benefit for few-shot learning, domain-generalization and robustness against adversarial attacks and label noise. In future, we hope to explore other variants of OPL including its adaptation to unsupervised representation learning.

References

[1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, volume 31, 2018. 7

[2] Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? *Advances in Neural Information Processing Systems*, 31:4261–4271, 2018. 2

[3] Fabio M. Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7

[4] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390*, 2020. 2

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020. 7

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269, 2019. 6

[7] Jiankang Deng, J. Guo, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019. 1, 2, 3, 6

[8] Guillaume Desjardins, Karen Simonyan, Razvan Pascanu, and Koray Kavukcuoglu. Natural neural networks. *arXiv preprint arXiv:1507.00210*, 2015. 2

[9] Qi Dou, Daniel C. Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 5

[10] Qi Dou, Daniel C. Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 7

[11] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Selecting relevant features from a universal representation for few-shot classification. *arXiv preprint arXiv:2003.09338*, 2020. 7, 8

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2017. 8

[13] Aritra Ghosh, H. Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017. 6

[14] Micah Goldblum, Steven Reich, Liam Fowl, Renkun Ni, Valeriia Cherepanova, and Tom Goldstein. Unraveling meta-learning: Understanding feature representations for few-shot tasks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3607–3616. PMLR, 13–18 Jul 2020. 2, 5, 12

[15] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 1

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006. 1, 2

[17] Munawar Hayat, Salman Khan, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Gaussian affinity for max-margin class imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6469–6479, 2019. 2, 3

[18] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019. 7

[19] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, 2020. 2, 6, 7

[20] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 103–112, 2019. 2

[21] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017. 2

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 6

[24] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[25] Eugene Lee and Chen-Yi Lee. Neuralscale: Efficient scaling of neurons for resource-constrained deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1478–1487, 2020. 12

[26] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 8

[27] José Lezama, Qiang Qiu, Pablo Musé, and Guillermo Sapiro. Ole: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8109–8118, 2018. 2, 6

[28] Aoxue Li, Weiran Huang, X. Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12573–12581, 2020. 2, 8

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 7

[30] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 2, 6

[31] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *ECCV*, 2020. 2, 8

[32] Weiyang Liu, Y. Wen, Zhiding Yu, Ming Li, B. Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, 2017. 1, 2, 3, 6

[33] Weiyang Liu, Y. Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. *International Conference on Machine Learning*, 2016. 1, 2, 3, 6

[34] Weiyang Liu, Yan-Ming Zhang, Xingguo Li, Zhiding Yu, Bo Dai, Tuo Zhao, and Le Song. Deep hyperspherical learning. In *Advances in Neural Information Processing Systems*, pages 3953–3963, 2017. 2

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 7

[36] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468)*, pages 41–48, 1999. 5

[37] Dmytro Mishkin and Jiri Matas. All you need is a good init. *arXiv preprint arXiv:1511.06422*, 2015. 2

[38] C. Qi and F. Su. Contrastive-center loss for deep neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2851–2855, 2017. 1, 2

[39] Shafin Rahman, Salman Khan, and Nick Barnes. Polarity loss for zero-shot object detection. *arXiv preprint arXiv:1811.08982*, 2018. 2

[40] Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Mubarak Shah. Self-supervised knowledge distillation for few-shot learning. *https://arxiv.org/abs/2006.09785*, 2020. 8

[41] A. Ravichandran, R. Bhotika, and S. Soatto. Few-shot learning with embedded class models and shot-free meta training. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 8

[42] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7957–7968. Curran Associates, Inc., 2019. 7, 8

[43] Pau Rodríguez, Jordi Gonzalez, Guillem Cucurull, Josep M Gonfaus, and Xavier Roca. Regularizing cnns with locally constrained decorrelations. *arXiv preprint arXiv:1611.01967*, 2016. 2

[44] Serim Ryou, Seong-Gyun Jeong, and Pietro Perona. Anchor loss: Modulating loss scale based on prediction difficulty. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 6

[45] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[46] J. Snell, Kevin Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017. 8

[47] Guolei Sun, Salman Khan, Wen Li, Hisham Cholakkal, Fahad Khan, and Luc Van Gool. Fixing localization errors to improve image classification. *ECCV*, 2020. 2, 6

[48] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3800–3808, 2017. 2

[49] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 8

[50] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 7

[51] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020. 2, 7, 8, 12, 13

[52] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. *http://arxiv.org/abs/1903.03096*, abs/1903.03096, 2019. 7

[53] Weitao Wan, Yuanyi Zhong, Tianpeng Li, and Jiansheng Chen. Rethinking feature distribution for loss functions in image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6

[54] Feng Wang, Xiang Xiang, Jian Cheng, and A. Yuille. Normface: L2 hypersphere embedding for face verification. *Proceedings of the 25th ACM international conference on Multimedia*, 2017. 1

[55] H. Wang, Yitong Wang, Z. Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wenyu Liu. Cosface: Large margin cosine loss for deep face recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 1, 2, 3, 6

[56] Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X Yu. Orthogonal convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[57] Yitong Wang, Dihong Gong, Zheng Zhou, Xing Ji, Hao Wang, Z. Li, W. Liu, and T. Zhang. Orthogonal deep features decomposition for age-invariant face recognition. In *ECCV*, 2018. 2

[58] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020. 7

[59] Y. Wen, Kaipeng Zhang, Z. Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 1, 2, 3, 6

[60] Di Xie, Jiang Xiong, and Shiliang Pu. All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6176–6185, 2017. 2

[61] Han-Jia Ye, Hexiang Hu, D. Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8805–8814, 2020. 2

[62] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 2

[63] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. 2017. 6

[64] Xiao Zhang, Rui Zhao, Yu Qiao, and Hongsheng Li. Rbf-softmax: Learning deep representative prototypes with radial basis function softmax. In *ECCV*, 2020. 2, 6, 12

[65] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 2018. 6, 7

[66] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016. 2

Supplementary: Orthogonal Projection Loss

In this supplementary document we include:

- Additional comparisons with the baseline on the MNIST dataset (Appendix A.1).
- OPL performance with scalable neural architecture method [25] (Appendix A.2).
- Robustness of OPL against noise in the input images (Appendix A.3).
- Visualization of classification results (Appendix B).

Appendix A. Experimentation

Here, we present results of experiments conducted using OPL on a set of additional settings.

Appendix A.1. Digit Classification (MNIST)

We conduct experiments on the MNIST dataset integrating OPL over a CE baseline. We use a 4-layer convolutional neural network with 32-dimensional feature embedding (after a global average pool operation) following the experimental setup in [64]. Our results are reported in Table 1. Additionally, we conduct experiments appending a fully-connected layer to reduce the feature dimensionality to 2 for generating better visualizations on behaviour of OPL in feature-space (presented in Fig. 1 in main article).

Method	1st	2nd	3rd	Avg
CE (baseline)	99.28%	99.27%	99.25%	99.27%
CE+OPL (ours)	99.58%	99.56%	99.61%	99.58%

Table 1: **Results on MNIST:** OPL obtains improvements over the CE baseline on MNIST dataset. Each experiment is replicated thrice and the average across runs is also reported.

Appendix A.2. Scalable Architectures

We consider the recent Neural Architecture Scaling approach proposed in [25] and plug-in our OPL on top of it to study our scalability. Refer Table 2 for the results.

Method	Backbone	Baseline[25]	[25] + OPL
NeuralScale [25]	ResNet18	77.59%	**77.81%**
NeuralScale [25]	VGG11	67.42%	**67.69%**

Table 2: **Additional results on CIFAR-100:** Performance improvements integrating OPL into small-scalable backbones for classification. Reported values are top-1 classification accuracies.

Appendix A.3. Robustness to Noise: FSL

We have already established through empirical evidence how OPL improves performance for few-shot learning tasks as well as robustness to adversarial examples present during evaluation. We now explore the more challenging task of exploring robustness to input sample noise in a FSL setting (similar to the one in Appendix B.1). The base training is conducted with no noise present in training data. During evaluation, the support and query set images are corrupted with random Gaussian noise of varying standard deviation (referred to as σ). This can be considered a domain shift on top of unseen novel classes during evaluation. The features learned with OPL during base training exhibit better robustness to such input corruptions in this FSL setting. We report these results in Table 3. The experiments conducted followed the method in [51] integrated with OPL.

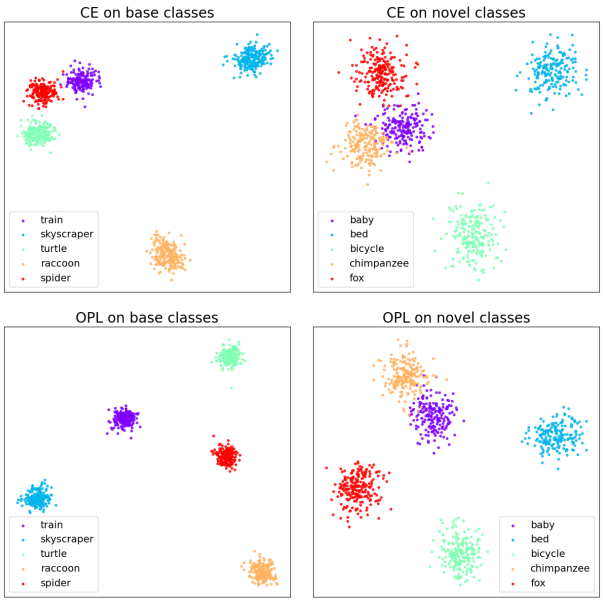


Figure 1: **LDA visualization for CE vs OPL in FSL setting:** Training with OPL increases separation of features in both base and novel classes when applied in a few-shot learning setting. LDA has been used following the insights in [14].

Appendix B. Visualization

In this section, we present additional visualizations exploring various aspects of OPL and its performance.

Method	Noise	Cifar:1shot	Cifar:5shot	Mini:1shot	Mini:5shot	Tier:1shot	Tier:5shot
RFS [51]	($\sigma = 0.1$)	63.30±0.39	80.36±0.28	55.98±0.37	74.46±0.27	66.54±0.43	82.92±0.29
RFS[51] + OPL	($\sigma = 0.1$)	**65.42±0.40**	**81.41±0.30**	**56.21±0.36**	73.20±0.29	**66.60±0.41**	**83.21±0.29**
RFS [51]	($\sigma = 0.05$)	68.32±0.38	84.34±0.27	60.22±0.36	77.45±0.27	68.65±0.41	83.12±0.27
RFS[51] + OPL	($\sigma = 0.05$)	**71.05±0.41**	**84.46±0.28**	**61.70±0.37**	**77.59±0.27**	**69.60±0.40**	**84.50±0.29**

Table 3: **Additional FSL Experiments:** We explore the robustness of models to noise (random Gaussian noise of varying standard deviation is added to input images) in FSL setting. Models trained with our proposed OPL loss are significantly more robust compared to the cross-entropy only baseline in [51].

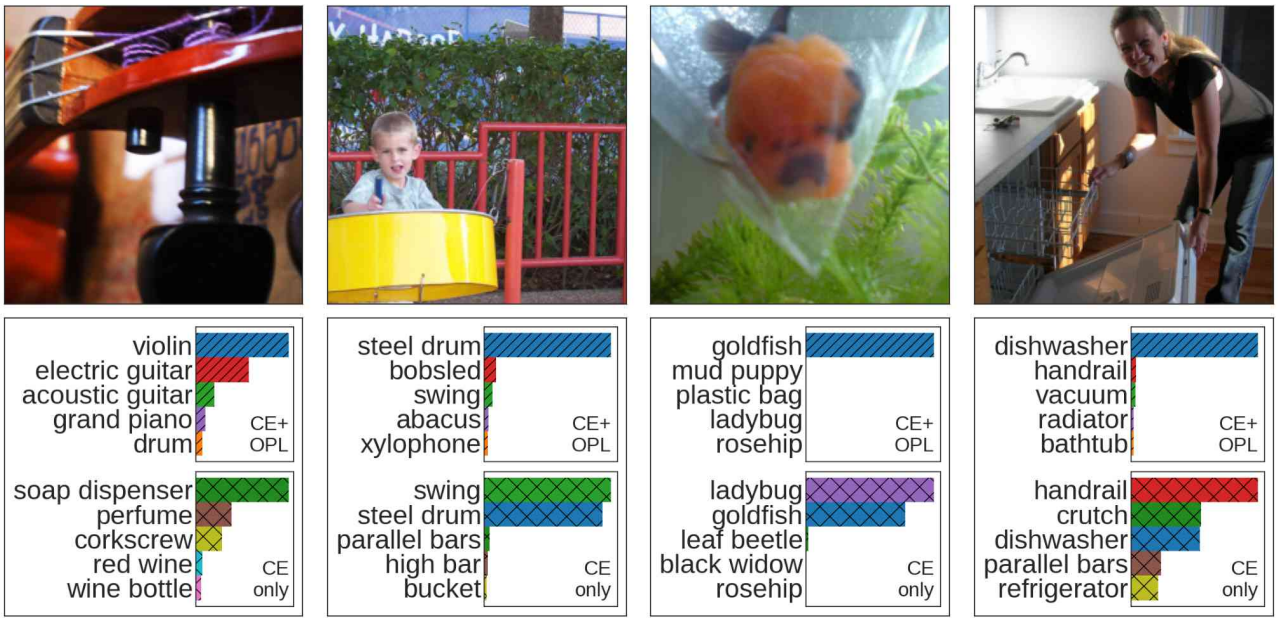


Figure 2: **Visualization of Images:** we show images where OPL predicts the correct but CE fails.

Appendix B.1. Class Embeddings

Consider a few-shot learning setting, where a model trained in a fully-supervised manner (referred to as base model / base training) on a set of selected classes which contains training labels (referred to as base classes) is later evaluated on a set of unseen classes (referred to as novel classes). The sets of base and novel classes are disjoint. The evaluation protocol would involve episodic iterations, where in each step a small set of labelled samples from the novel classes (referred to as support set) is available during inference, and an another set of those same novel classes (referred to as query set) is available for calculating the accuracy metrics.

Given how our proposed loss is already able to explicitly enforce constraints on the feature space during base training, we want to examine if the additional discriminative nature endowed on the features by OPL is aware of higher level semantics. To evaluate this, we explore the more challenging task of inter-class separation and intra-class clustering of novel classes which are unseen during the base training. We train a model following the approach in [51] integrating OPL, and visualize the separation of different class features for both base and novel classes in Fig. 1.

Appendix B.2. ImageNet Examples

We further explore the performance of our model (CE+OPL) trained on ImageNet (model used for experiments presented in Table 2 of main paper) by examining the failure cases of the baseline model that were improved upon when adding OPL. Visualizations for some randomly selected such cases are illustrated in Fig. 4 and Fig. 2.

Appendix B.3. Block Matrix

We defined the overall objective of OPL as a minimization of the expected inter-class orthogonality (refer Eq. 8 of main paper) and conducted empirical analysis using models training using our proposed loss function against a CE only baseline (illustrated in Fig. 4 of main paper). In this section, we conduct additional analysis on those block-matrices to

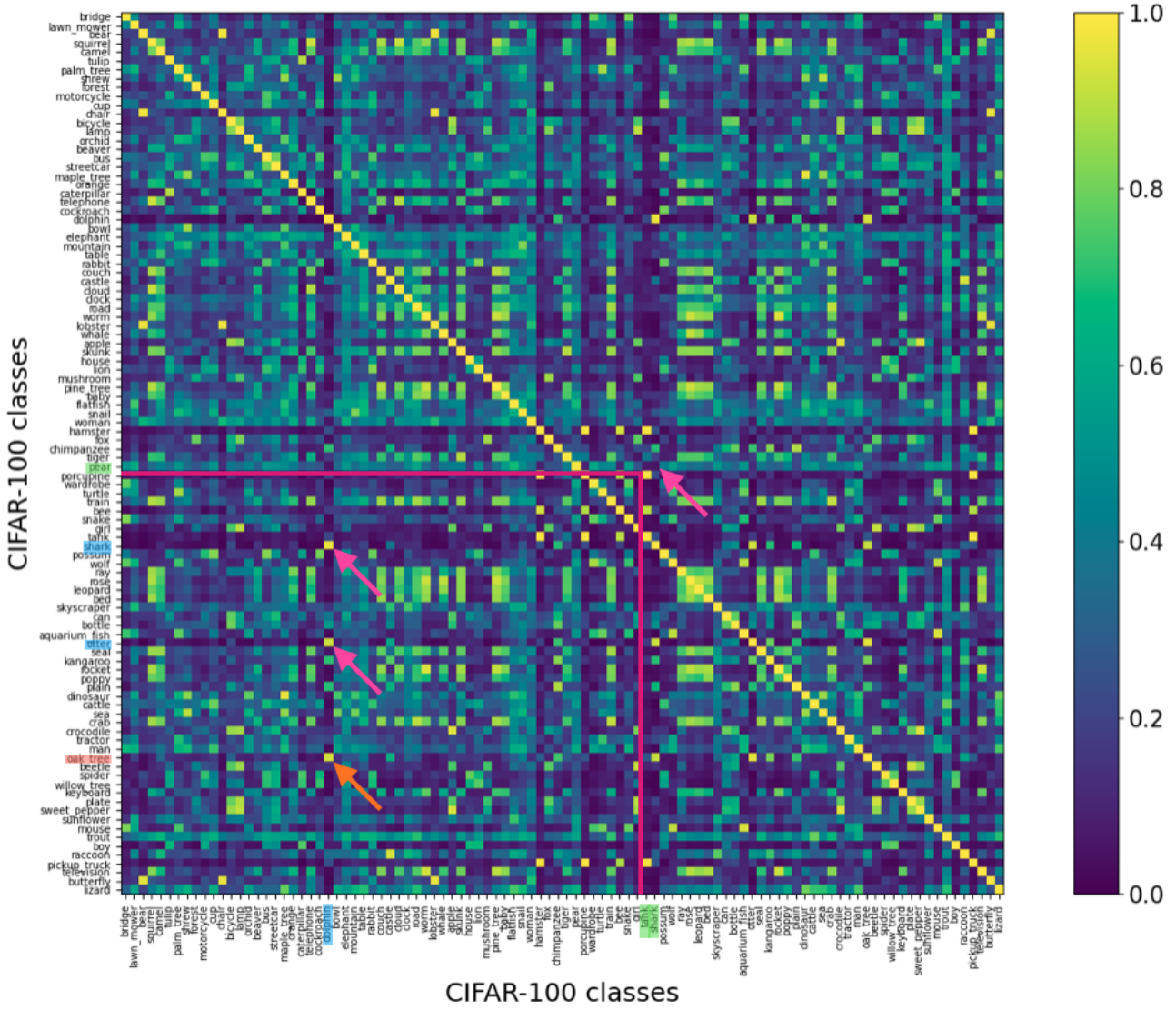


Figure 3: **Orthogonality Visualization:** The diagram (enlarged and elaborated version of Fig 4b in main paper) visualizes the cosine similarity between each pair of per-class feature vectors extracted from an OPL trained ResNet-56 for the CIFAR-100 test-set. Each per-class feature vector is calculated averaging over the features of all samples belonging to that class within the test-set. We analyse the relationships for two randomly selected classes, *dolphin* and *pear*. Consider the similarity of the dolphin class column (label highlighted in blue). In general, it has low similarity with the other classes, except in 3 instances. Two of those, *shark* and *otter* (pink arrows) align with our heuristics on similarity of those categories. The similarity to *oak tree* category can be attributed to some correlation present within the test-set images of these two classes (*e.g.*, both contain large blue portions - ocean for *dolphin* and sky for *oak tree*). Now, consider pear (label highlighted in green), which has an average similarity to most other classes except two: *tank* and *shark* (labels highlighted in green / *tank* in CIFAR-100 is the military vehicle). These two classes have relatively lower similarity with the *pear* class as seen from the diagram (pink lines and pink arrow) which again aligns with our intuition about the relationships between these categories. Overall, we note that the orthogonality constraints we enforce on feature space through OPL allows room for learning hidden inter-class relationships which can be interpreted meaningfully, in comparison to the same relationships for the CE baseline.

further understand the outcomes of our orthogonality constraints on the learned feature space. It is interesting to note that while OPL enforces a higher degree of orthogonality between the average class vectors, it does not naively push everything to be orthogonal. We note that this allows any hidden knowledge learned during the training process (information not captured in the labels explicitly) to remain captured within the features. The results of the experiments

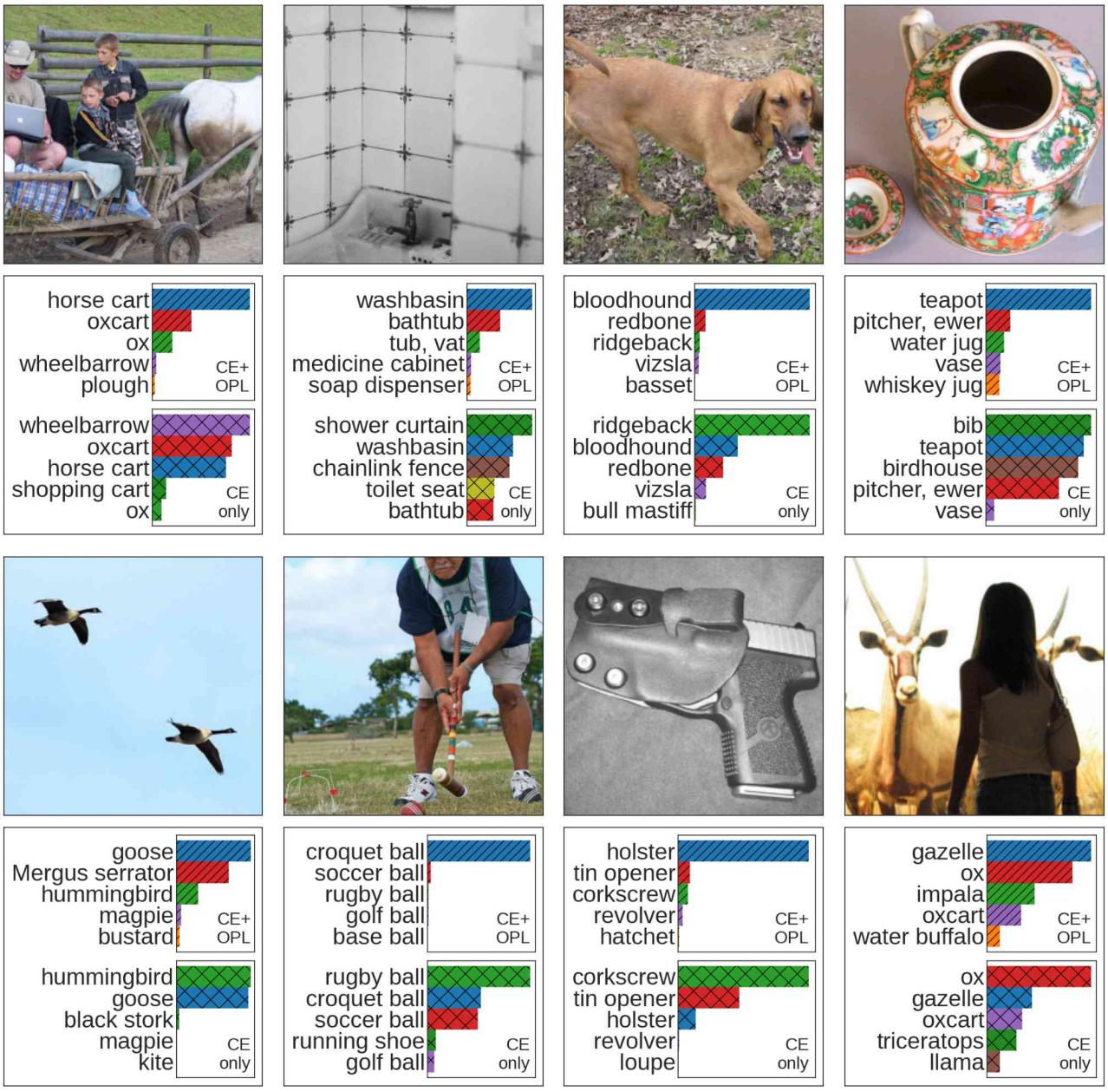


Figure 4: **Visualization of Classification Results:** We show some examples of images where OPL predicts the correct class but CE fails.

conducted on this are illustrated in Fig. 3.