Supplementary: Orthogonal Projection Loss

In this supplementary document we include:

- Additional experimentation on ImageNet dataset (Appendix A.1).
- Additional comparisons with the baseline on MNIST dataset (Appendix A.2).
- OPL performance with scalable neural architecture method [24] (Appendix A.3).
- Robustness of OPL against noise in the input images (Appendix A.4).
- Visualization of learned representations and classification results (Appendix B).

Appendix A. Experimentation

We present results for additional experiments conducted using OPL on a variety of settings.

Appendix A.1. Additional ImageNet results

We include additional experimentation over a state-of-the-art baseline for the Imagenet classification task in Table 1.

Method	Top-1 (%)	Top-5 (%)
ResNet50-D [18]	78.31→**79.26**	94.09→**94.62**

Table 1: Increase in accuracy on ImageNet val. set with OPL.

Appendix A.2. MNIST results

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 [63]. Our results are reported in Table 2. 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 2: **Results on MNIST:** OPL obtains improvements over the CE baseline on MNIST dataset. Each experiment is replicated thrice and the average across runs is additionally reported.

Appendix A.3. Scalable Architectures

We run experiments using smallar, scalable deep neural network architectures where we plug-in OPL on top of their methodology. These experiments are based on the approach followed in [24]. Refer Table 3.

Method	Backbone	Baseline[24]	[24] + OPL
NeuralScale [24]	ResNet18	77.59%	**77.81%**
NeuralScale [24]	VGG11	67.42%	**67.69%**

Table 3: **Additional results on CIFAR-100:** Performance improvements integrating OPL into small-scalable backbones for classification. Reported values are top-1 classification accuracies.

Appendix A.4. 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 one in Appendix B). 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 4. The experiments conducted followed the method in [49] integrated with OPL.

Quantitative results highlighting these performance improvements are presented in Table 4.

Appendix A.5. Additional Ablation

We conduct further ablations on few-shot learning and label noise tasks in Table 5 and Table 6. We also evaluate the sub-components of OPL as well as an alternate variant of covariance from [61] in Table 7.

Appendix B. Visualization

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

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

Method	Noise	Cifar:1shot	Cifar:5shot	Mini:1shot	Mini:5shot	Tier:1shot	Tier:5shot
RFS [49]	($\sigma = 0.1$)	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[49] + 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 [49]	($\sigma = 0.05$)	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[49] + 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 4: **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 [49].

Param	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 2.0$
γ=2.0	64.41	64.52	62.03	58.55	54.82
γ=1.0	64.36	64.73	65.11	61.33	58.68
γ=0.5	63.47	63.76	64.83	63.02	59.85

Table 5: FSL results (1-shot) on held-out val. set of CIFAR-FS.

Param	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 2.0$
γ=2.0	55.11	54.19	*6.07	*3.66	*3.59
γ=1.0	54.77	52.75	56.70	56.96	60.02
γ=0.5	53.78	52.45	50.48	49.49	33.99

Table 6: Label noise results: Accuracy on CIFAR-100 val. set.

Model	CE only	CE + s	CE + d	CE + OPL	CE + cov [61]
Top-1	72.40	71.06	70.34	**73.52**	72.99
Top-5	92.68	91.65	91.42	**93.07**	92.84

Table 7: The sub-components of OPL (s & d) individually do not create the desired effect of simultaneous clustering and separation. OPL also outperforms minimizing mini-batch covariance matrix.

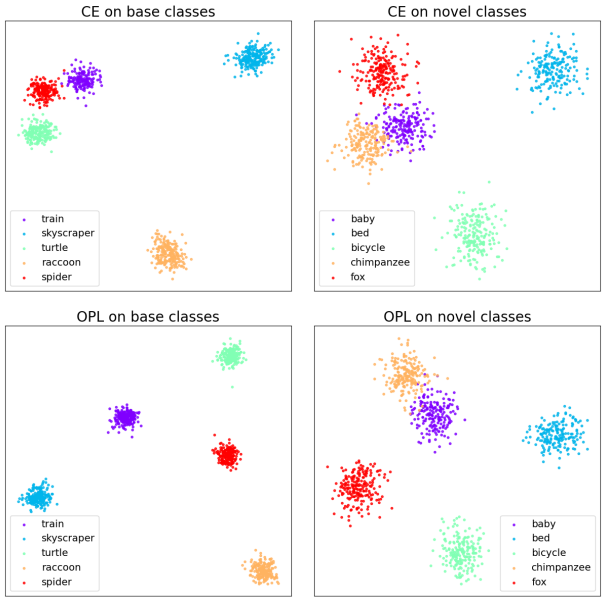

Figure 1: **LDA visualization for CE vs OPL in FSL setting:** Training with OPL increases separation of features in both base and novel classes when applied in a few-shot learning setting. LDA has been used following the insights in [14].

contain 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 for during inference fine-tuning, and another set of those same novel classes (referred to as query set) is available for calculating the accuracy metrics. While there is room for fine-tuning of the base model during inference, we note that the feature space of that model is mostly defined during the base training.

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 clsutering of novel classes which are unseen during the base training. We train a model following the approach in [49] 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 by examining the failure cases of the baseline model that were improved upon when adding OPL. These results are illustrated in Fig. 3.

Appendix B.3. Block Matrix

We defined the overall objective of OPL as a minimization of the expected inter-class orthogonality (refer Eq. 8) 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 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

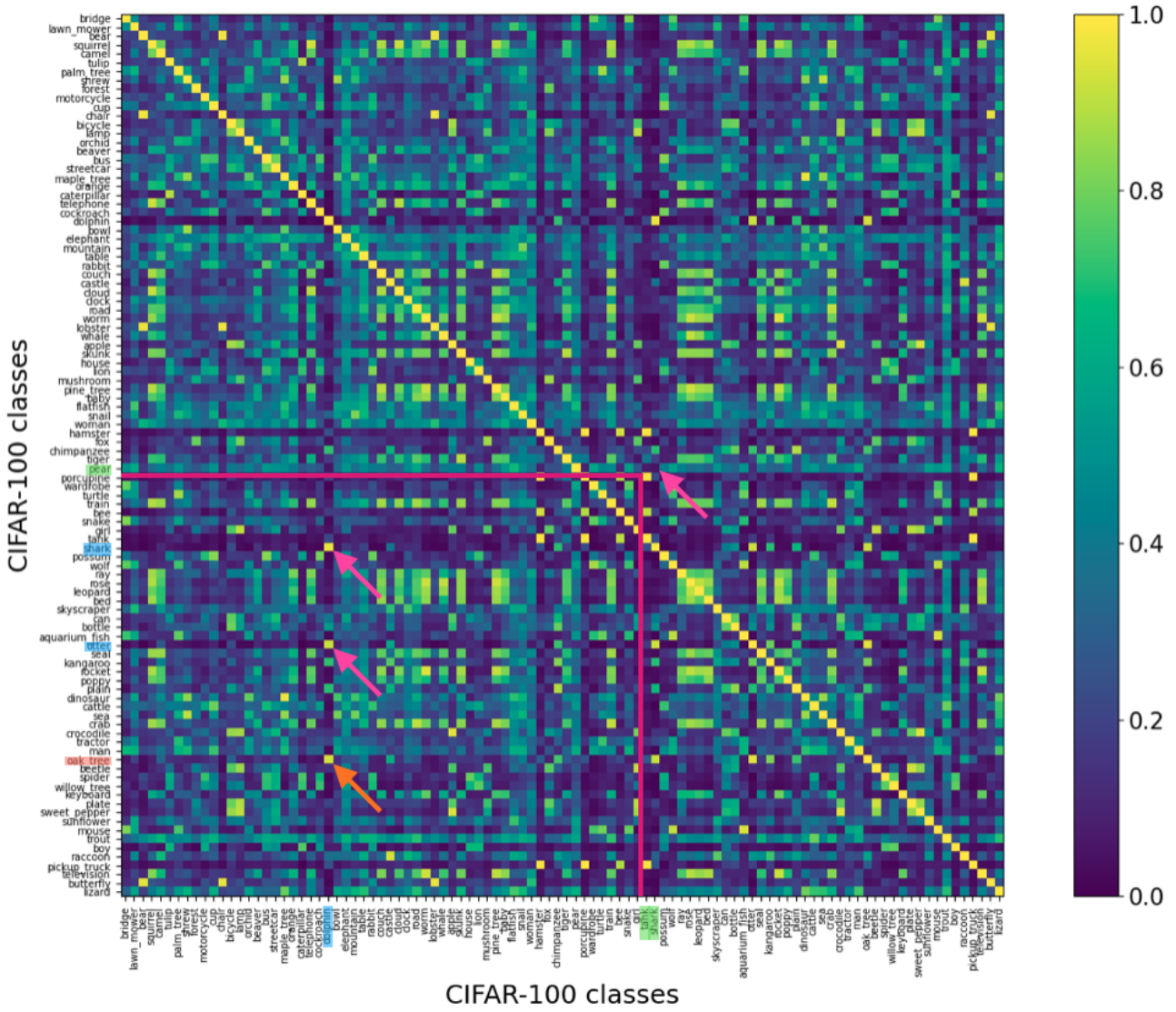

Figure 2: **Orthogonality Visualization:** The diagram (enlarged version of Fig 4b in main paper) visualizes the cosine similarity between each pair of per-class feature vectors extracted from an OPL trained ResNet-56 for the CIFAR-100 test-set. Each per-class feature vector is calculated averaging over the features of all samples belonging to that class within the test-set. We analyse the relatioships 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 outcomes of the constraints we enforce on feature space through OPL can be interpreted meaningfully to a greater extent in comparison to the same relationships for the CE baseline.

be orthogonal. We note that this allows any hidden knowledge learned during the training process (information not captured in the labels explicitly) to remain within the features. The results of the experiments conducted on this are illustrated in Fig. 2.

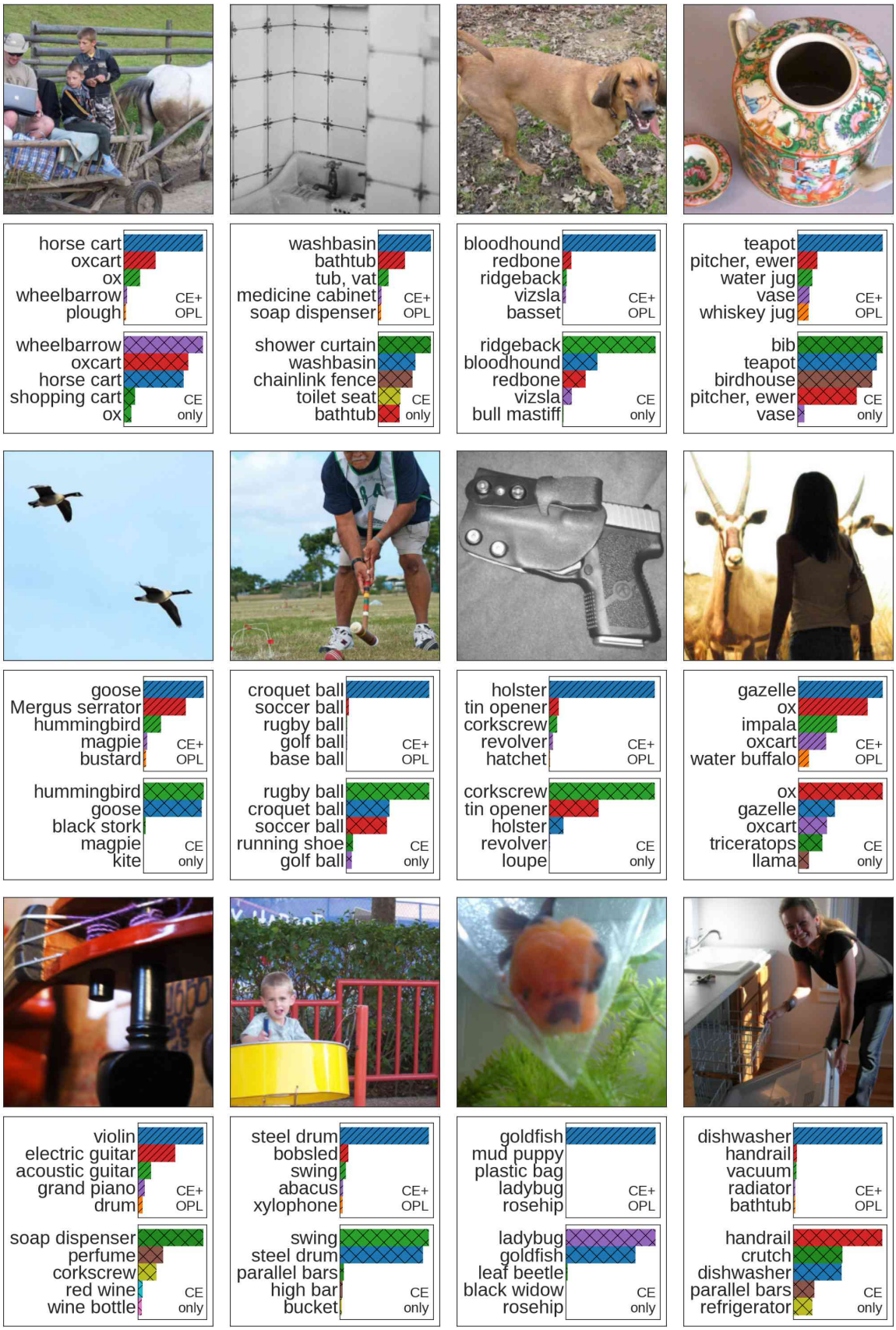

Figure 3: **Visualization of Images:** we show images where OPL predicts the correct but CE fails.