# Novel Ensemble Diversification Methods for Open-Set Scenarios Supplementary Material

Miriam Farber
Amazon
Haifa, Israel
mffarber@amazon.com

Roman Goldenberg
Google Research *
Haifa, Israel
romanfg@google.com

George Leifman
Google Research *
Haifa, Israel
gleifman@google.com

Gal Novich
Amazon
Haifa, Israel
ganovich@amazon.com

The supplementary material is divided into two sections. In section 1 we provide an in depth discussion on the embedding spaces generated by the ensemble models, as well as on several additional theoretical aspects of the proposed diversification approaches. In section 2, we present additional experimental results and discuss implementation details, which include hard negative mining, structure of the training batches, and the full details on hyper-parameters used for training. These details, together with those mentioned in the main paper, allow full reproducibility of our results.

## 1. Extended Discussion

### 1.1. Exploring Diversity via Embedding Space

To better understand the impact of the proposed diversification methods on ensembles, we examine the internal representation (embedding $e(x)$ in the main paper) learned by ensemble models. Throughout this section we analyze two-model ensembles trained on CIFAR-10 [4] using OSCRL and FDL methods with varying diversification parameter $\alpha$.

#### 1.1.1 OSCRL - in-distribution data embedding:

Three ensembles (two models) were trained by OSCRL with a varying $\alpha \in \{0, 0.1, 0.2\}$. Sup. Fig. 1 visualizes the learned embedding spaces for the three pairs of models. The embeddings of in-distribution data samples are projected by the PCA [9] onto 2D and color-coded according to their class (10 classes of CIFAR-10).

Qualitatively, one can see the changing degree of similarity between the embedding spaces of ensemble models as $\alpha$ grows. For $\alpha = 0$, the embedding spaces of the two models are practically a mirror of each other. For $\alpha = 0.1$, the spaces are of a similar topology, with some color swaps. For $\alpha = 0.2$, models produce embeddings with different topology, with many color swaps.

For quantitative analysis, we use the following Mahalanobis-like metric to measure the distance between two classes in the embedding space of the training data: given two classes $a, b$, we calculate the mean embedding (centroids) for each class $\bar{x}_a, \bar{x}_b$, and intra-class co-variance matrices $S_a$ and $S_b$, the distance between the two classes is computed as

$$d(\bar{x}_a, \bar{x}_b) = \sqrt{(\bar{x}_a - \bar{x}_b)(S_a + S_b)^{-1}(\bar{x}_a - \bar{x}_b)}$$

We compute the distances between the centroids of all $K$=10 classes, yielding a pair-wise distance matrices $D_1$ and $D_2$ for the two ensemble models. Considering $S = |D_1 - D_2|$ we evaluate the change in centroid relations between models, and by extension the similarity between the two embeddings, by measuring the mean of the matrix S. Sup. Fig. 2 presents the similarity matrices $S$ over various $\alpha$. As expected, the changes are more pronounced as $\alpha$ grows: the mean value of the matrix $S$ changes from $0.17$ to $0.67$ and $0.82$, for $\alpha = 0, 0.1, 0.2$, respectively.

#### 1.1.2 FDL - distilled data embedding:

In this section we visualize the embedding space of ensemble models trained using the Feature-Based Diversification Loss (FDL). In FDL, we start with an existing model $f_1$. For each image $x$, we distil its features and encode them as a corresponding $x_{f_1}$ image. Then, we train a model $f_2$ in a way that encourages it to disregard features learned by $f_1$. The $f_2$ model is trained to correctly classify $x$'s by mapping them into a separable, discriminative embedding space. At the same time, $f_2$ is being encouraged to map $x_{f_1}$'s into an inseparable/indiscriminative subspace of the embedding.

Sup. Fig. 3 provides a visualization of the $f_1$ and $f_2$ embedding spaces. The visualization is done by UMAP [5] transform (as implemented by [6]) with 20 neighbors, using the Euclidean metric. We used this clustering method here as it emphasizes the separation between classes and the compactness of each class.

*This work was conducted under Amazon.

Figure 1. **Higher $\alpha$ leads to shifts in closed-set embedding:** Embedding spaces of three ensembles trained by OSCRL with $\alpha \in \{0, 0.1, 0.2\}$ (two models per ensemble). Points are color-coded according to their class. Centroids of each class are marked with a circle. Compare the class (color) neighbourhood relations for the two models of each ensemble. For $\alpha = 0$, the neighbourhood graph is identical for the two models. As $\alpha$ grows, the neighbours are not preserved between the two models.

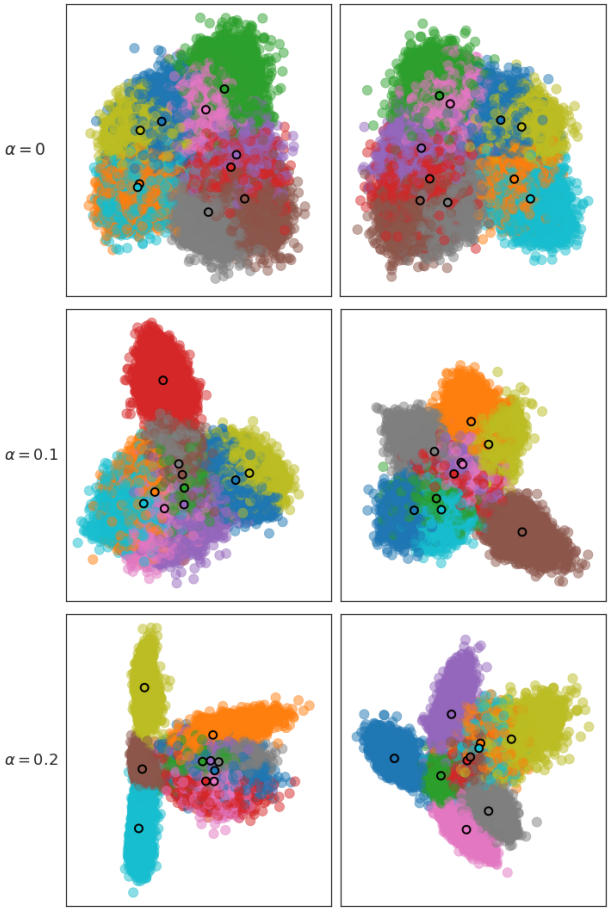



Figure 2. **Higher $\alpha$ leads to greater difference between ensemble models embeddings:** $S = |D_1 - D_2|$ matrices (where $D_i$ is the matrix of pair-wise distances between centroids of $K = 10$ classes for model $i = 1, 2$) for $\alpha \in \{0, 0.1, 0.2\}$ (left to right). All color-values are set on the same linear scale spanning over the extremal values in the figure [0, 2.1].

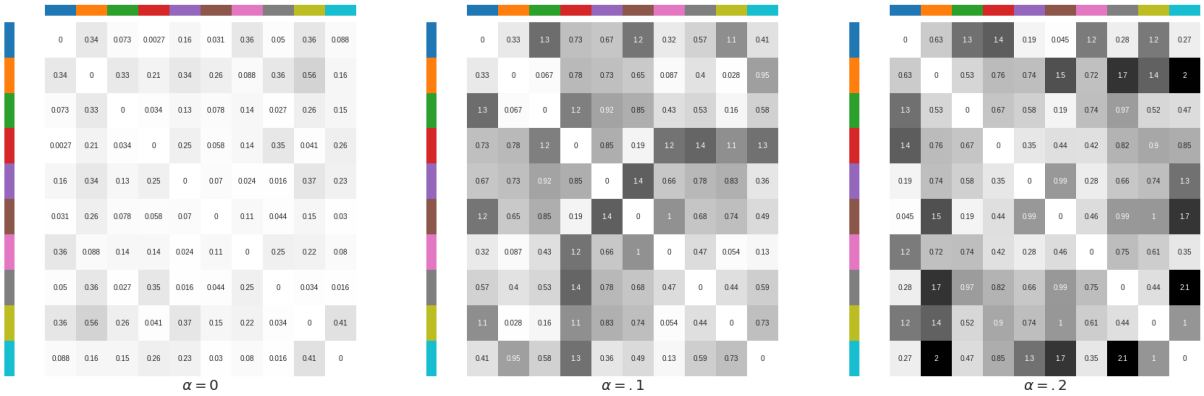
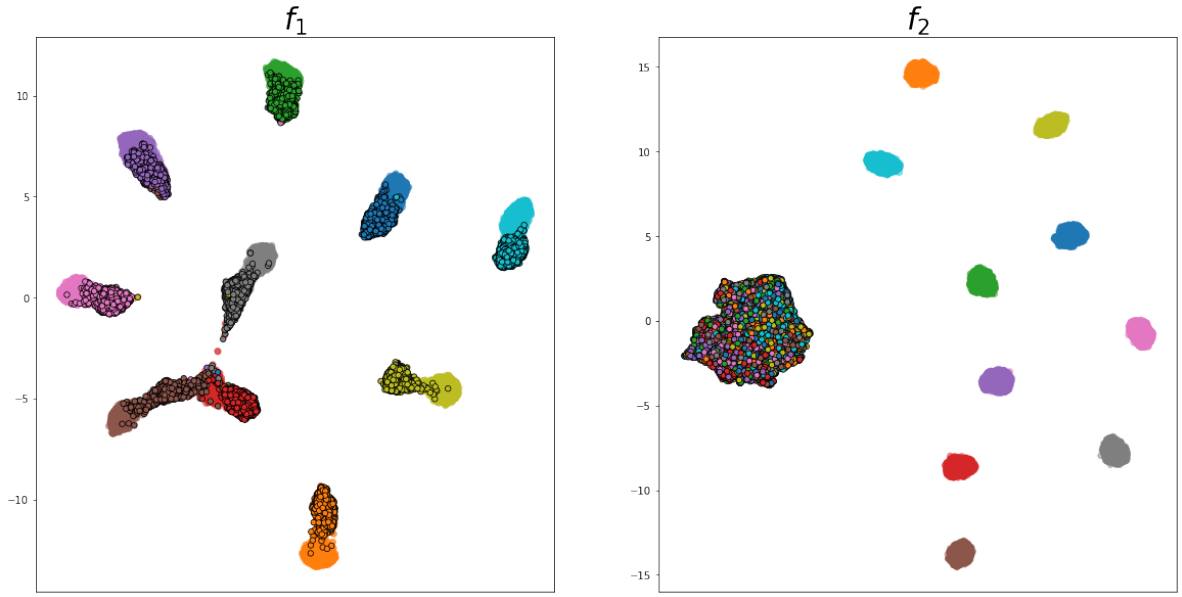


Figure 3. **FDL embedding:** A UMAP transform of the $f_1$ (left) and $f_2$ (right) embeddings. All $f_1$-distilled images (circles with black boundary) of all classes are clustered in a compact, inseparable subspace by $f_2$.



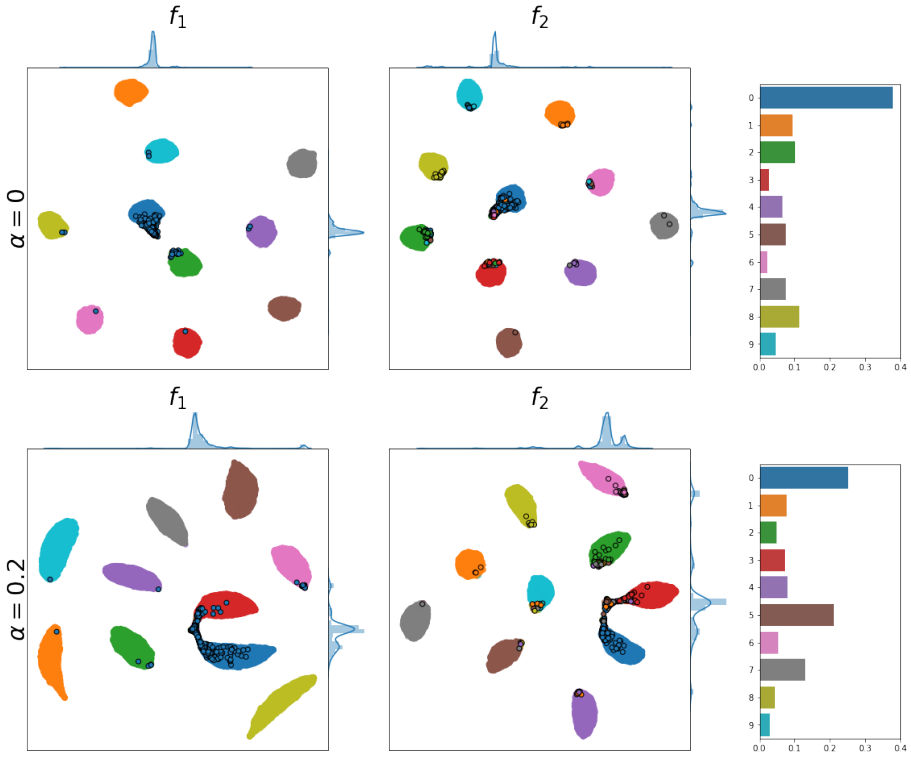

The training set points are presented by circles with no boundary, color-coded by their ground-truth class. In addition, points corresponding to embeddings of distilled images are shown as colored circles with a black edge. The distilled images have been confirmed to be classified to the same class as their source by $f_1$, and are color-coded accordingly. The left image shows the embedding space of $f_1$. As UMAP reduces the dimensionality, while preserving global distances, the distilled image points are indeed close to their matching class in the $f_1$ embedding. The right image depicts the embedding of the same samples by $f_2$. Remarkably, the embeddings of distilled images of all classes are now clustered in an inseparable, compact subspace.

We further substantiate our claim quantitatively. Let $x$ be a sample of class $q(x)$ and $x_{f1}$ be its distilled image. Let us denote by $c_{(q(x),f)}$ the centroid of the class $q(x)$ in the emebdding space induced by model $f$. We measure the Mahalanobis distance between a distilled image to its class centroid for the two ensemble models $f_1$ and $f_2$: $d(f_1(x_{f_1}), c_{(q(x),f_1)})$, $d(f_2(x_{f_1}), c_{(q(x),f_2)})$. We then average these distances over all distilled images in the training set to yield $\hat{d}_{f1}$ and $\hat{d}_{f2}$. We repeat this experiment twice for ensembles trained with $\alpha = 0$ (independently trained models) and $\alpha = 0.1$. For $\alpha = 0$, $\hat{d}_{f1}$ and $\hat{d}_{f2}$ are approximately the same (around $4.82$). For $\alpha = 0.1$, we get $\hat{d}_{f1} = 4.82$. This time, $\hat{d}_{f2} = 8.03$, almost 2 times higher than $\hat{d}_{f1}$.

#### 1.1.3 OSCRL - out-of-distribution (OOD) samples embedding:

We claim in the paper that the open set diversification loss encourages ensemble models to disagree on inputs belonging to unknown classes, resulting in accuracy increase on open sets. In this section we provide some visual insights to substantiate this claim. In Sup. Fig. 4, we depict the embeddings obtained by two ensembles, trained with $\alpha = 0$ (top row) and $\alpha = 0.2$ (bottom row). Each ensemble includes two models- $f_1$ (left column in the figure) and $f_2$ (middle column). We use UMAP here as well [5]. In-distribution samples from the 10 classes are presented for both $f1$ and $f2$ as color-coded filled circles with no boundary. In addition, for $f1$, we depict the embeddings of out-of-distribution samples (marked by dark blue circles with black boundary) that $f_1$ classifies as belonging to class 0 (dark blue). In the middle column, we visualize the embeddings of the same samples as encoded by $f_2$. For each such sample $x$, we color the embedding of $f_2(x)$ according to the color of the class predicted by $f_2$, and mark it with black edge to differentiate it from the in-distribution embeddings. Finally, in the right column, we present a histogram showing the distribution of the classes predicted by $f_2$ for these samples. We can clearly see the reduction of the dark blue bar when using $\alpha = 0.2$ as opposed to $\alpha = 0$, demonstrating that the open set diversification loss indeed managed to decrease the agreement rate of the ensemble models on the outliers mapped to class 0 via $f_1$ (from $37.8\%$ with $\alpha = 0$ to $25.1\%$ with $\alpha = 0.2$).

Figure 4. **Out-of-distribution (OOD) samples in the embedding space:** Embedding spaces of two ensembles (each ensemble consists of two models $f_1$ and $f_2$) trained by OSCRL with $\alpha = 0$ (top row) and $\alpha = 0.2$ (bottom row). Out-of-distribution (OOD) samples mapped to the blue class by $f_1$ are marked by circles with black boundary. The histogram on the right shows the distribution of classes predicted by $f_2$ for those OOD samples. One can see that $f_2$ of the $\alpha = 0$ ensemble agrees with $f_1$ on the blue class on almost 40% of the samples. For $\alpha = 0.2$, on the other hand, the agreement rate of $f_2$ and $f_1$ is much lower (about 30%), and in about 20% of the cases $f_2$ classified OOD samples as "brown"




## 1.2. Diversification via explainability

In this section we show how the diversity induced by the OSCRL method encourages ensemble models to concentrate on different aspects/regions of the recognized object. As in the previous section, we use the CIFAR-10 dataset for demonstration and an ensemble of three models.

To figure out which image regions are more important for a given model to perform a specific task, we use the popular Grad-CAM selvaraju2017grad technique. Grad-CAM uses the class-specific gradient information flowing into the final convolutional layer of a CNN to produce a coarse localization map of the important regions in the image.

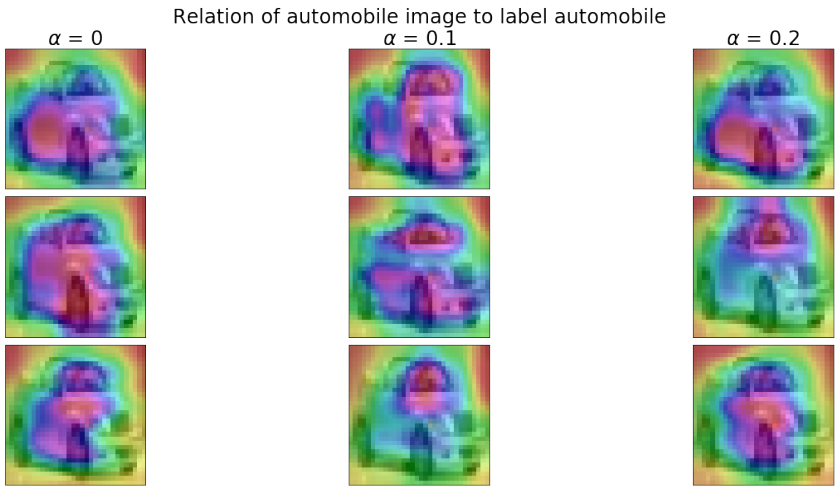
Sup. Fig. 5 presents the importance maps for three ensemble models (the three rows) trained using the OSCRL with three different $\alpha$ values (the three columns). The importance maps are with regard to the task of classification of a car image as a car (correct classification was achieved in all presented cases). See how the focus of attention of the three ensemble models diverges to different car parts with higher $\alpha$.

We further experiment with Counterfactual Explanations, as described in the Grad-CAM paper [10]. In Sup. Fig. 6, we present the importance map for the classification of a horse image as a horse, along with the counterfactual information encouraging it not to be classified as a deer. The image was correctly classified as a horse by all models. We can see the completely identical attention maps for the horse classification over all models. However, counterfactual maps are getting more and more diverse with higher $\alpha$. This visualization further confirms that OSCRL loss encourages ensemble models to disagree on inputs belonging to unknown classes.

## 1.3. Open-set diversification loss term reduction

In the paper we defined the model similarity metric as the "average pair-wise correlation between model predictions on the wrong classes" (see Eq. 1). In OSCRL, this metric is an explicit term in the loss function. Naturally, this leads to a reduction in the value of this similarity metric along the

Figure 5. **Higher $\alpha$ yields diverse car recognition ensembles:** The activation of the final CNN layer is presented for each model of the ensemble (the three rows). Highly important areas are marked in magenta. Ensembles of the same $\alpha$ are grouped by column. Visibly, lower $\alpha$ produces consistent mappings, while higher $\alpha$ produces different explanations for the classification of the image.




OSCRL-based training process.

$$L_{\text{Corr}}(x) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \text{Corr}(f_i \setminus q(x), f_j \setminus q(x)) \quad (1)$$

Surprisingly, even when this metric does not appear explicitly in the loss function, such as for Joint-based training or for the baseline (two models trained independently), we observed that it's value, when measured after the training is finalized, is lower than it's value at the beginning of the training process (see Sup. Fig. 7). The fact that we see this phenomenon even in the baseline training may indicate that the randomness in the training process results in models that utilize different features, hence leading to a reduction in the similarity metric. With Joint training, this phenomenon is even stronger, indicating that ensemble of models trained via Joint training benefits from diversification (as was also discussed in the context of equation 4 in the paper).

### 1.4. Accuracy as a function of $\gamma$ (percentage of outliers)

Fig. 3 in the main paper shows the graphs of ensemble accuracy ($acc$) for open-set classification as a function of $\gamma$. Peculiarly, for CIFAR-10 the $acc(\gamma)$ functions are decreasing, while for CIFAR-100 they are increasing. In order to gain a more in-depth understanding of the asymptotic behaviour of ensemble models as a function of $\gamma$, we would like to derive a formula for the slope of these graphs. For an open-set classification problem with $K$ classes, the accuracy $acc$ depends on the accuracy of the system on each one of the $K+1$ classes ($K$ "in-distribution" classes and one outlier class), and on $\gamma$. Let $acc_{in}$ and $acc_{out}$ be the overall system accuracy on the first $K$ (in-distribution) classes, and the OOD class respectively. These quantities represent the proportion of correctly classified classes (out of the entire set of relevant classes). Finally, denote by $X_{in}$ the amount of in-distribution samples. Then the amount of OOD samples is $\gamma \times X_{in}$. Thus:

$$acc = \frac{acc_{in} \times X_{in} + acc_{out} \times \gamma \times X_{in}}{X_{in} + \gamma \times X_{in}} \quad (2)$$

We can now measure the impact of change in $\gamma$ on the open-set accuracy, while keeping $X_{in}, acc_{in}$ and $acc_{out}$ constant:

$$\frac{\partial acc}{\partial \gamma} = \frac{acc_{out} - acc_{in}}{(1+\gamma)^2} \quad (3)$$

Eq. 3 can explain the above-mentioned $acc(\gamma)$ slope differences. Indeed, $acc_{in}$ is expected to be lower for CIFAR-100 (100-class problem is harder than 10-class). On the other hand, $acc_{out}$ is expected to be higher for CIFAR-100, as it is less likely that ensemble models agree on the same class (out of 100) for an outlier. Hence, for CIFAR-100 it is more likely getting $acc_{out} > acc_{in}$, and, thus, an increasing $acc(\gamma)$. This also demonstrates the advantage of open-set tailored diversification methods, that excel even more when the proportion of OOD data is larger (due to relatively high $acc_{out}$ compared to other methods).

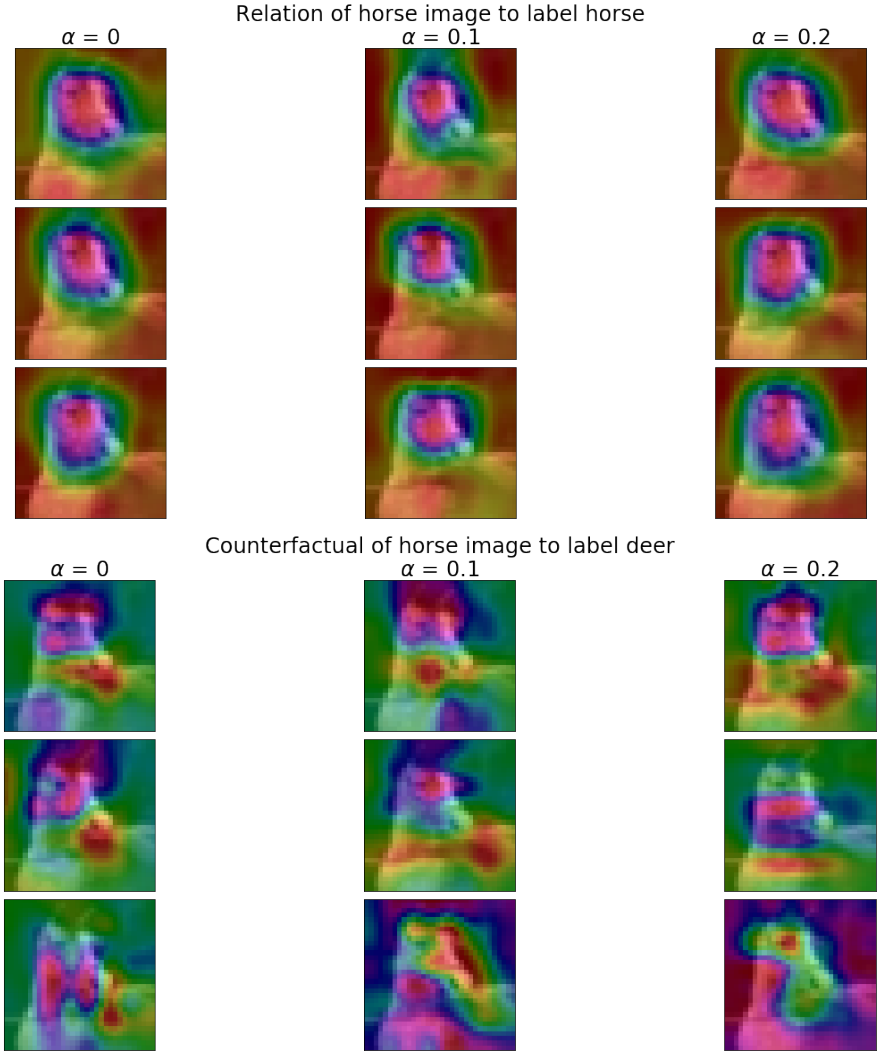
## 2. Implementation details and analysis

### 2.1. Implementation details - OSCRL and FDL

#### 2.1.1 OSCRL Implementation details

As mentioned in the paper, the open-set ensemble diversification methods we propose are motivated by the ques-

Figure 6. **Higher $\alpha$ yields diverse counterfactual recognition ensembles:** See caption of Supplementary Figure 5 for technical explanation. In the upper part of the figure we show the Grad-CAM selvaraju2017grad generated importance map for the classification of a horse sample as the horse class, which in this case is very consistent for all models. In the lower part of the figure, the counterfactual information for classification as a deer is presented. Specifically, for $\alpha = 0$ two of the models focus on the lack of antlers, the other on jaw shape. The $\alpha = 0.2$ ensemble models partition the image into distinctive and complementary areas.




tion "how to train an ensemble to disagree on unknown data/outliers if this data is unavailable during training?"

In OSCRL-based training, we use the "wrong class" probabilities generated by the models on valid input as a proxy for model output on outliers. That is, we train on known-class data, but request the inter-model disagreement on wrong class probabilities only. To reiterate, the open-set correlation reduction loss (OSCRL) is defined as:

$$OSCRL(x) = \frac{1-\alpha}{n}\sum_{i=1}^{n} L_{CE}(q(x), f_i(x)) + \alpha L_{\text{Corr}}(x),$$

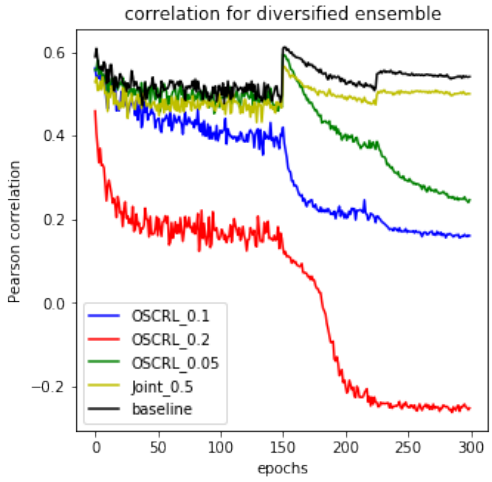
with $L_{\text{Corr}}(x)$ as defined in Eq. 1 and

$$f_i \backslash q(x) = (f_i^1(x), \ldots, f_i^{q(x)-1}(x), f_i^{q(x)+1}(x), \ldots, f_i^n(x)).$$

In the above formulas, $q(x)$ represents the ground-truth class to which $x$ belongs. As we have shown in the main paper, OSCRL outperforms the competing methods on the vast majority of the classification tasks. On recognition tasks, it is usually ranked $2^{nd}$ after FDL. In this section, we would like to expand on OSCRL training process. As can be seen in Eq. 1, the loss function contains terms of the form

$$\text{Corr}(f_i \setminus q(x), f_j \setminus q(x)).$$

Figure 7. Similarity metric (Eq. 1) during the training process for the baseline, Joint, and OSCRL-based training. The subscript near labels is the value of $\alpha$ used for training.




If $f_i^{q(x)}(x)$ is very large compared to the second largest component in $f_i(x)$ (and same for $f_j$), the elements of $f_i \setminus q(x)$ and $f_j \setminus q(x)$ are small, making the task of minimizing the correlation $\mathrm{Corr}(f_i \setminus q(x), f_j \setminus q(x))$ very easy. Thus when training with OSCRL, we employ hard negative mining procedure, including the terms on the RHS of Eq. 1 only when

$$\max_{t \neq q(x)} (f_i^{q(x)}(x) - f_i^t(x))$$

is below a predefined threshold for at least one of the ensemble models $f_i$. We used the threshold 0.9 for experiments with datasets that include small number of classes (MNIST, SVHN, CIFAR-10), and a threshold of 0.2 for experiments that included datasets with large number of classes (CIFAR-100 and the recognition datasets). The rationale for choosing the threshold was to eliminate cases in which the correct class got a significantly higher probability compared to the wrong classes. In practice, we have seen that the models are not too sensitive to the thresholds, and the importance of the mining process was in removing loss terms that are very easy to minimize, hence not forcing the model to improve on harder cases.

#### 2.1.2 FDL Implementation details

The feature-based diversification training process is described in the main paper and formally defined by Equations (8) and (9). To reproduce our results it is important to treat batch normalization carefully. Note that the loss function in Equation (9) is based both on regular and distilled images. These images are taken from two different distributions and therefore they differ in their statistics. Since batch normalization uses running statistics, it is important to make sure the training batches consist of even mixture of regular and distilled images.

### 2.2. Extended experimental results

This section provides complete results and implementation details for the experiments discussed in the main paper.

#### 2.2.1 Classification

Table 1 presents extended experimental results on CIFAR-10 and CIFAR-100 classification tasks (see Table 1 in the main paper). The hyperparameters used for training are taken from [11]. We also used their GitHub repository [1], replacing only the loss functions with those we present in the paper. For each one of the models and datasets, we use SGD with batch size of 64. The initial learning rate of 0.1 is decreased by a factor of 10 at epochs 150 and 225, with momentum 0.9. Results are averaged over 3 training trials of random parameters initialization. The experiments were performed on p3.2xlarge AWS instances.

As can be seen from Table 1, on CIFAR-10, OSCRL outperforms the rest of the methods on a wide range of $\alpha$'s. For CIFAR-100, OSCRL and FDL dominate for the smaller network (DN-64-6), while for mid-size network OSCRL is the best, followed by FDL and Joint, with NCL being the last. For the large network (DN-100-12), NCL outperforms the rest. Note that for NCL, using $\alpha = 0.1$ (and above) deteriorates the performance significantly. We believe that this is due to the nature of cross-entropy, which is sensitive and tends to "explode" when all probability vector components are small (hence, a low value for $\alpha$ is required).

#### 2.2.2 Adversarial Attacks

An OOD input can be contextually interpreted as an adversarial attack. Prior research shows that ensemble diversification translates well to general adversarial robustness [7, 1]. Following [7], we conducted an evaluation of OSCRL under the Fast Gradient Signed Method (FGSM), Basic Iterative Method, and Projected Gradient Descent attacks. We trained ensembles of DenseNets 82-8-8 on CIFAR-10 using the OSCRL loss, with $\alpha \in [0, 0.7]$. OSCRL consistently outperforms the baseline ($\alpha = 0$) for all tested attacks with significant accuracy increase, while the accuracy degradation for the no-attack scenario is minimal (see Table 2).

We use ensembles of 3 DenseNets 82-8-8 models, trained on CIFAR-10, using OSCRL, in the same way as for the classification task. The adversarial attacks were implemented with a Pytorch version of CleverHans [8], using the

[1] https://github.com/grey-area/modular-loss-experiments

Table 1. **Results on CIFAR** Accuracy of ensembles trained using 5 different diversification approaches: independently trained models - baseline, Joint training, NCL, OSCRL and FDL, with $\gamma = 100\%$. Top table - CIFAR-100. Bottom table - CIFAR-10. Rows correspond to different network architectures - three types of DenseNets. The best results within each row are in green, the best results within each diversification approach are in blue.

| CIFAR-100 | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| architecture | baseline | Joint | | | NCL | | | | FDL | | | | OSCRL | | | |
| $\alpha$= | - | .1 | .25 | .5 | .02 | .05 | .07 | .1 | .02 | .05 | .1 | .2 | .02 | .05 | .1 | .2 |
| DN-64-6 | 60.9 | 61.8 | 60.9 | 62.8 | 61.2 | 59.8 | 62.5 | 50.0 | 63.5 | 63.6 | 62.8 | **64.2** | 62.0 | 62.2 | 62.4 | **64.2** |
| DN-82-8 | 65.1 | 65.6 | 67.3 | 65.1 | 64.2 | 64.6 | 66.3 | 50.0 | 67.1 | 67.0 | 66.8 | 65.9 | 65.5 | 65.8 | 66.1 | **69.4** |
| DN-100-12 | 65.4 | 68.3 | 68.0 | 69.6 | **69.9** | 68.3 | 63.3 | 50.0 | 66.3 | 67.7 | 66.0 | 66.5 | 67.8 | 69.0 | 68.1 | 68.7 |

| CIFAR-10 | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| architecture | baseline | Joint | | | NCL | | FDL | | | | OSCRL | | | |
| $\alpha$= | - | .25 | .5 | .75 | .02 | .05 | .02 | .05 | .1 | .2 | .05 | .1 | .15 | .2 |
| DN-64-6 | 73.8 | 73.5 | 74.0 | 75.8 | 72.7 | 73.7 | 72.7 | 72.6 | 73.0 | 73.5 | 75.4 | 75.8 | **77.9** | 77.5 |
| DN-82-8 | 72.3 | 71.2 | 73.7 | 75.6 | 72.3 | 72.3 | 72.4 | 72.3 | 72.3 | 72.4 | 76.9 | **77.0** | 76.6 | 75.8 |
| DN-100-12 | 71.2 | 71.6 | 73.3 | 75.5 | 71.1 | 72.6 | 71.9 | 71.6 | 71.5 | 72.2 | 76.9 | 78.2 | 78.2 | **78.4** |

Table 2. **OSCRL improves robustness to adversarial attacks** Accuracy of ensembles on CIFAR-10 under various attacks (rows). The ensembles include 3 DenseNets(82-8-8) models trained with OSCRL, using different $\alpha$'s (columns). Best results in green.

| | | $\alpha$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Attack | Param | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 |
| FGSM | $\epsilon = 0.02$ | 69.0 | 73.4 | 74.3 | 75.8 | **76.9** | 76.5 | 76.4 | 76.5 |
| | $\epsilon = 0.04$ | 48.5 | 55.9 | 59.2 | 60.8 | 62.9 | **63.2** | 62.4 | 63.5 |
| BIM | $\epsilon = 0.01$ | 75.0 | 75.0 | 76.9 | 77.0 | **77.9** | 77.4 | 77.2 | 77.1 |
| | $\epsilon = 0.02$ | 46.4 | 50.1 | 54.6 | 56.9 | 58.1 | 57.8 | **58.8** | 58.6 |
| PGD | $\epsilon = 0.01$ | 79.9 | 79.3 | 81.3 | 80.9 | **81.8** | 81.1 | 81.0 | 80.9 |
| | $\epsilon = 0.02$ | 55.5 | 57.1 | 61.2 | 62.7 | 64.1 | 63.4 | 64.3 | **64.4** |
| No Attack | | **95.5** | 95.4 | 95.1 | 94.7 | 94.5 | 94.3 | 94.4 | 93.6 |

parameters from [7]. OSCRL ensembles consistently improve the robustness to adversarial attacks, while negligibly degrading the accuracy on "no attack", even for a relatively high baseline ($\alpha = 0$) performance. Notice the optimal diversification factor $\alpha$ increases with the severity of attacks.

### 2.2.3 Person Re-Identification

To perform a fair comparison, for all compared diversification methods, we used the same ResNet50 models trained using an SGD optimizer with the following parameters: batch-size of 32, number of epochs 60, starting learning rate of 0.05 with the decay by a factor of 10 every 20 epochs.

Table 3 shows TTR for different FTRs, for various ensemble diversification methods. It can be seen that for all the FTR values, the FDL yields the best result, outperforming the baseline by up to $30\%$. The OSCRL is the $2^{nd}$ best approach. Note that FDL outperforms all other methods with a single $\alpha$ value (0.05), which yields the best accuracy for all FTRs. For other methods, the optimal $\alpha$ value depends on the FTR target.

In addition we define the *non-target agreement rate*, $RANK - 1_{non}$, which measures the chance that, when removing the correct identity from the gallery, the ensemble models will agree on a wrong top identity. In general, the lower this rate, the lower the chances to accept wrong non-target probe, regardless of the individual match scores.

### 2.2.4 Face Recognition

All the ensembles were comprised of IR-SE50 nets (a combination of IR-50 and SENet [2]) and trained with a Stochastic Gradient Descent optimizer. We used the following parameters: batch-size of 240, 20 epochs, starting learning rate 0.1 with weight decay of 5e-4 on parameters that are not part of batch normalization. All models were trained on p3.8xlarge AWS instances (which include 4 Tesla V100-SXM2 GPUs).

Table 4 shows TTR for different FTRs, for various ensemble diversification methods. FDL and OSCRL outper-

Table 3. **Re-ID on Market-1501 [12] benchmark:** Re-identification accuracy (TTR) at different FTR targets for the 5 ensemble diversification approaches. The best results are in green, $2^{nd}$ best in blue.

| method | Baseline | NCL | | | Joint | | | FDL | OSCRL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha =$ | 0 | 0.01 | 0.05 | 0.1 | 0.2 | 0.5% | 0.9% | 0.05 | 0.2 | 0.5 | 0.7 |
| FTR=1% | 16.9% | 20.1% | 21.1% | 20.2% | 19.4% | 19.2% | 19.6% | **22.2%** | 20.3% | 20.0% | 18.9% |
| FTR=10% | 56.3% | 54.7% | 53.8% | 55.5% | 53.8% | 55.6% | 53.9% | **61.3%** | 56.9% | 56.6% | 60.9% |
| FTR=20% | 71.6% | 69.0% | 68.9% | 71.2% | 69.7% | 69.8% | 70.6% | **76.8%** | 72.9% | 72.9% | 75.6% |
| FTR=30% | 78.3% | 76.4% | 76.5% | 78.3% | 76.6% | 77.7% | 78.0% | **83.5%** | 80.1% | 79.5% | 80.0% |
| $RANK\text{-}1_{non}$ | 55.8% | 32.1% | 31.0% | 39.2% | 52.6% | 51.9% | 51.6% | 42.1% | 40.1% | 32.3% | **27.4%** |

Table 4. **Face recognition on LFW [3] benchmark:** Face recognition accuracy at FTR=$\{0.5\%, 1\%\}$ for 5 ensemble diversification approaches (independently trained models (baseline), NCL, Joint training, FDL, and OSCRL) for various $\alpha$. Cells with the best result for each approach are colored in blue, and the best overall result in green.

| method | baseline | NCL | Joint | | | | | FDL | | | OSCRL | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha=$ | 0 | .05 | .1 | .25 | .5 | .75 | .9 | .01 | .05 | .1 | .1 | .25 | .5 | .75 | .9 |
| TTR@FTR=1% | 98.26 | 98.42 | 98.27 | 98.07 | 98.17 | 98.07 | 98.17 | **98.57** | 98.48 | 98.49 | 98.50 | 98.47 | 98.46 | 97.96 | 97.48 |
| TTR@FTR=.5% | 83.33 | 85.71 | 84.11 | 83.38 | 83.67 | 83.38 | 83.70 | 85.04 | 84.64 | 84.48 | 83.06 | 85.61 | 84.57 | 83.98 | **86.41** |
| $RANK-1$ | 99.83 | 99.82 | 99.83 | **99.84** | 99.81 | **99.84** | 99.81 | 99.81 | 99.82 | 99.80 | 99.83 | 99.83 | 99.81 | 99.79 | 99.71 |

form other methods and the relatively high baseline. The $RANK-1$ score is the chance for the ensemble to agree on the correct top identity when the target exists in gallery. All models seem to achieve comparable performance. Additionally, the OSCRL model with high $\alpha$ shows some trade-off between TTR@FTR%5 and $RANK-1$ values.

## References


[1] Mahdieh Abbasi, Arezoo Rajabi, Christian Gagné, and Rakesh B Bobba. Toward adversarial robustness by diversity in an ensemble of specialized deep neural networks. In *Canadian Conference on Artificial Intelligence*, pages 1–14. Springer, 2020.

[2] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[3] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, 2008.

[4] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[5] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, Feb. 2018.

[6] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[7] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. *arXiv preprint arXiv:1901.08846*, 2019.

[8] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018.

[9] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[10] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[11] Andrew M Webb, Charles Reynolds, Dan-Andrei Iliescu, Henry Reeve, Mikel Luján, and Gavin Brown. Joint training of neural network ensembles. *arXiv preprint arXiv:1902.04422*, 2019.

[12] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.