LORD: Leveraging Open-Set Recognition with Unknown Data

Tobias Koch[*, †], Christian Riess[†], and Thomas Köhler[*]
[*]e.solutions GmbH, Erlangen, Germany
[†]Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
{tobias.koch, thomas.koehler}@esolutions.de, christian.riess@fau.de

Abstract

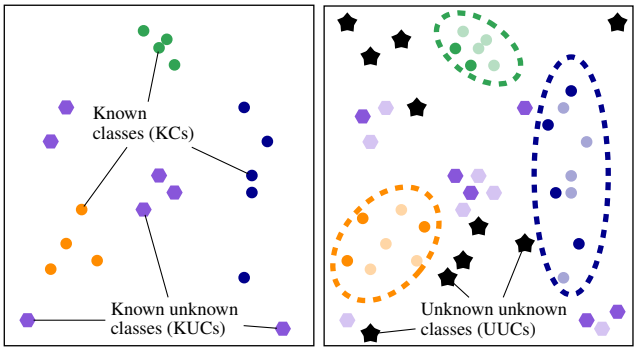
Handling entirely unknown data is a challenge for any deployed classifier. Classification models are typically trained on a static pre-defined dataset and are kept in the dark for the open unassigned feature space. As a result, they struggle to deal with out-of-distribution data during inference. Addressing this task on the class-level is termed open-set recognition (OSR). However, most OSR methods are inherently limited, as they train closed-set classifiers and only adapt the downstream predictions to OSR.

This work presents LORD, a framework to Leverage Open-set Recognition by exploiting unknown Data. *LORD explicitly models open space during classifier training and provides a systematic evaluation for such approaches. We identify three model-agnostic training strategies that exploit background data and applied them to well-established classifiers. Due to LORD's extensive evaluation protocol, we consistently demonstrate improved recognition of unknown data. The benchmarks facilitate in-depth analysis across various requirement levels. To mitigate dependency on extensive and costly background datasets, we explore mixup as an off-the-shelf data generation technique. Our experiments highlight mixup's effectiveness as a substitute for background datasets. Lightweight constraints on mixup synthesis further improve OSR performance.*

1. Introduction

Most classification algorithms are designed for *closed-set* environments, where all classes are known prior to deployment of the classifier. However, real-world applications may expose classifiers to unseen classes. Recognizing that an input belongs to an unseen class constitutes the open-set recognition (OSR) task. Scheirer *et al*. [50] distinguish two subtasks: 1) *Recognizing* samples as known or unknown. 2) *Classify* knowns to its specific class. An example is face recognition of few known subjects among many others [20].

We distinguish three types of classes [51] as depicted in Fig. 1: 1) *Known classes (KCs)* are uniquely labeled examples detected *and* classified. They are essential in both the training and test set. 2) *Known unknown classes (KUCs)* is background data and comprises samples not necessarily grouped into meaningful categories, but assumed to be unrelated to the KCs. They are part of the training and test set. We differentiate *genuine* KUCs, a dataset subset, from *synthesized* KUCs. 3) *Unknown unknown classes (UUCs)* are unseen during training. They are only part of the test set.



(a) Training data (opaque). (b) Test data (opaque) with training data (shaded).

Figure 1. Overview of data types in open-set recognition. The training set in (a) includes known classes (KCs) and known unknown classes (KUCs) (⬢). The trained classifiers in (b) model decision boundaries for KCs as dashed ellipses. KUCs correlate with the training set's KUCs, exhibiting higher identifiability in comparison to the unknown unknown classes (UUCs) (★), which can exist anywhere in the feature space.

In recent years, OSR has gained increasing attention [17, 49, 61]. Most approaches adopt *closed-set training with open-set inference*, as exemplified in various methods, including distance-based classifiers [4, 29, 38, 47], Support Vector Machines (SVMs) [3, 19, 27, 51], Extreme Value Machines (EVMs) [23, 30, 48, 58], losses and calibrations for Deep Neural Networks (DNNs) [5, 39, 54], uncertainty quantification [6, 34, 35], novelty detection [7], and autoencoder architectures [41, 42, 63]. For inference, a threshold is introduced to either reject or classify a test sample based on a confidence value [56].

Accepted at ICCVW 2023.

Let $\mathcal{T}_K = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^T$ be a labeled training dataset with features $\boldsymbol{x} \in \mathbb{R}^D$ and class labels $y \in \mathcal{C}_K$, where $\mathcal{C}_K$ denotes the set of KCs. Given $\mathcal{T}_K$, the goal is to infer a likelihood function $f(\boldsymbol{x}; \mathcal{T}_K) = \boldsymbol{z}$ with $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{C}_K|}$ that maps features to a confidence $z_y \in \mathbb{R}$ for each class in $\mathcal{C}_K$. For instance, $\boldsymbol{z}$ could describe normalized posterior probabilities or distances. Function f is used to decide for class $\hat{y}$ with the highest confidence: $\hat{y} = \arg\max_{\hat{y} \in \mathcal{C}_K} f(\boldsymbol{x}; \mathcal{T}_K)_{\hat{y}}$. A reject option converts this prediction into OSR according to:

$$g(\boldsymbol{x}, \delta; \mathcal{T}_K) = \begin{cases} \hat{y} & \text{if } \max_{\hat{y} \in \mathcal{C}_K} f(\boldsymbol{x}; \mathcal{T}_K)_{\hat{y}} > \delta \\ u & \text{otherwise} \end{cases}, \quad (1)$$

where δ denotes the decision threshold. However, training exclusively on a closed set can benefit the classification task but inherently limits the OSR capabilities.

To the best of our knowledge, only a few works consider background samples [13, 20, 43]. While genuine KUCs undoubtedly simplify OSR, they may not be available. Recent approaches show that synthesizing out-of-distribution (OOD) data with generative adversarial networks (GANs) can suffice for training DNNs [16, 40, 52]. However, GAN training can be unstable [31] and is designed for in-distribution data [18], necessitating additional methods to handle outliers [16]. The mentioned works offer valuable insights into end-to-end OSR with KUCs. However, most are trained in a K+1 fashion, where the background is modeled in a single class. This approach is limited by the background class's complexity [13]. An exception is the work by Günther *et al*. [20] that studies OSR via linear discriminant analysis and the EVM. We extend their work by exploiting KUCs in diverse ways and examine whether synthetic data is a suitable proxy for genuine KUCs.

Our framework LORD provides guidance for researchers to explicitly model open space during classifier training. To verify the efficacy of this modeled space on unknown recognition, a benchmark protocol is introduced. The protocol requires, besides KCs, access to genuine KUCs and UUCs. LORD distinguishes between two assessments: 1) biased assessment, involving training with genuine KUCs and testing with genuine KUCs and UUCs. 2) Unbiased assessment excludes KUCs during testing. This identifies optimal learning conditions for different models and application-specific requirements. In the context of face recognition, a biased application could be a system at an airport likely to encounter genuine KUCs, while an unbiased application might be the authentication on a smartphone.

LORD encompasses a comprehensive range of OSR metrics that facilitate in-depth analysis across various requirement levels. These levels may place a particular emphasis on secure open-set performance or prioritize strong closed-set performance, catering a more user-friendly behavior. In particular, this extensive evaluation highlights a fundamental trade-off between open- and closed-set performance across models.

To model open space, LORD deploys three model-agnostic training strategies on six classifiers spanning four distinct categories. These strategies improve biased OSR performance by up to 30 %, while unbiased performance generally demonstrates a more modest improvement.

Admitting the scarcity and cost of domain-specific background data in real-world applications, LORD offers a solution in the form of mixup as an off-the-shelf data generation technique. Mixup samples are convex combinations of distinct KCs and act as substitutes for genuine KUCs. While naïve mixups constitute the occupation problem, LORD proposes effective constraints to refine the generation process. Our experiments confirm mixup as an excellent substitute for genuine KUCs. This approach proves beneficial in assessing models before investing in resource-intensive collection of background data.

To summarize, LORD provides a systematic framework to assess the efficacy of background data exploitation to enhance classifiers within particular application contexts. When background data is not available, mixup poses an effective solution. If this proves beneficial, researchers might find it valuable to gather additional background data.

This work is organized as follows: Section 2 reviews related works. Section 3 introduces our OSR benchmarking protocol. Section 4 outlines the training strategies and their use for 6 OSR models. Section 5 explores mixup and the constrained generation. Section 6 summarizes the main findings and concludes this work.

2. Related work

We explore incorporating KUCs during training and reformulate Eq. (1) accordingly. Let $\mathcal{T} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^T$ be a *partially* labeled training dataset with class labels $y \in \mathcal{C}$, where $\mathcal{C} = \mathcal{C}_K \cup u$ and u indicates a missing label. Set $\mathcal{T}$ contains samples of KCs and KUCs, leading to a change in function g where $\mathcal{T}_K$ is replaced by $\mathcal{T}$, $h(\boldsymbol{x}, \delta; \mathcal{T}) = g(\boldsymbol{x}, \delta; \mathcal{T})$. The class decision remains unchanged, but the inclusion of KUCs impacts the likelihood function f, categorizing different approaches in OSR.

2.1. Semi-supervised model training

KUCs samples can also be considered as a special instance of unlabeled data. Exploiting both labeled and unlabeled data is a common focus in semi-supervised learning. Notable techniques in this field comprise manifold regularization [37] and the generation of pseudo-labels [2, 24], integrating unlabeled data into a supervised learning regime. Using unlabeled samples is considerably cheaper than labeled samples [59]. Jointly exploiting labeled and unlabeled data can substantially enhance the underlying models compared to purely supervised learning [65].

However, unlike OSR, semi-supervised learning tackles a complementary scenario: predictions cover only KCs. The approach assumes that unlabeled instances belong to one or more of the KCs, without considering the open space. We follow concepts similar to semi-supervised learning and incorporate unlabeled KUCs to improve OSR learning.

2.2. Open-set training with open-set inference

The vast body of previous works use unlabeled KUCs in a $K+1$ fashion, *i. e.*, K KCs and one unknown class [16, 40, 52]. In function h, $\mathcal{C}_K$ is substituted by $\mathcal{C}$, treating all samples with the label u as one large class. This allows the model to predict the background class without a threshold δ. However, we show that modeling unknowns in a single class can be suboptimal and is surpassed by alternative strategies.

Others use external datasets to define special loss functions to calibrate filters [43] or probabilities [13], particularly tailored to train DNNs. Günther *et al*. [20] take an initial step towards employing genuine KUCs with EVMs. They do not cluster KUCs but consider them to learn probabilistic models of KCs. This can improve the detection of UUCs compared to simple baselines like distance-based detection. These findings indicate the potential of leveraging KUCs to enhance the predictive power of OSR models.

In this paper, we generalize model-specific approaches into three different model-agnostic training strategies that exploit KUCs. Specifically, we extend multiple OSR models using these training strategies and determine the optimal strategy based on the underlying model and task.

2.3. Generating out-of-distribution samples

Recent works addressed the question of whether KUCs could be synthesized when external data is unavailable. For instance, Du *et al*. [14] model KUCs by estimating Gaussian distributions per KC and sampling background data near class boundaries. However, the Gaussian distribution assumption is simplistic, resulting in uniform restrictions of the decision boundary. Zhou *et al*. [67] regularizes DNN training using manifold mixup [57] to learn placeholders.

In-distribution data is synthesized through generative models [11, 18, 45, 53] or augmentation [10, 33, 44, 57, 64, 66]. Conversely, OOD data generation is more difficult. Ge *et al*. [16] propose a complex pipeline: 1) Train a DNN and conditional GAN using KCs. 2) Generate GAN samples and classify them with the DNN. 3) Store misclassified samples as KUCs. 4) Train another DNN in a $K+1$ way with KUCs as background class. This generation is guided by the assumption that regions where both models disagree do not belong to any KC. Unfortunately, achieving convergence of one GAN is challenging, leading to the training of several GANs on different class subsets. Also, control over the generated random samples is very limited, and the required number of KUCs remains an open question. Neal *et al*. [40] train an encoder-decoder GAN and a classifier. To generate KUCs, they alter KC samples in the latent space of the GAN to make the DNN uncertain about the prediction. The DNN is finetuned in a $K+1$ manner. Also here, there is a mutual dependence between the GAN and DNN and the generation process lacks control. Generating images is challenging due to their high dimensionality and the need to match the data context. Moreover, DNNs are vulnerable to small deviations as introduced by adversarial attacks [1, 15].

Kong and Ramanan [31] avoid mutual dependencies using DNN penultimate layer features to train a GAN. The GAN's discriminator detects unknowns, while the DNN classifies knowns. We adopt a similar two-stage approach, but unlike GANs operating in image space, we operate solely in the feature space. Specifically, we use manifold mixup [57] to generate KUCs, a lightweight alternative to complex GANs. However, manifold mixup may generate unwanted KUCs that overlap with a KC, referred to as *occupation problem*. To mitigate this, we introduce an effective filter to retain only meaningful KUCs.

3. How to benchmark with known unknowns?

Common OSR benchmarks [13, 40, 50] ignore KUCs for training and testing. In contrast, we extend the work of Günther *et al*. [20] to explore the impact of KUCs on model training. This serves as a reference under conditions when dataset-related background data is available.

3.1. Defining protocols with known unknowns

For our experiments, we define evaluation protocols using well-established datasets with varying characteristics.

CIFAR-100 [32]. This image dataset consists of 20 superclasses divided into 100 equally balanced subclasses. We split the superclasses into 12 KCs, 4 KUCs, and 4 UUCs. Adopting the subclass labels, this results in 60 KCs, 20 KUCs, and 20 UUCs. The training set contains 500 images per class, leading to a KUC to KC sample ratio of 0.33. For feature extraction, we use EfficientNet-B4 [55], pre-trained on ImageNet [12] and finetuned on the 60 KCs. The embedding dimension is 1792. Unless otherwise stated, we repeat all experiments 3 times with varying superclass-splits and report averaged results. While a pre-trained ImageNet feature extractor already saw all CIFAR-100 classes, this setup remains open-set for 3 reasons: 1) Image resolutions significantly differ. 2) Finetuning is subject to catastrophic forgetting [36, 46]. 3) We use the features to train a downstream classifier, which is never exposed to ImageNet.

Labeled Faces in the Wild (LFW) [25, 26]. For this face recognition dataset, we follow the protocol of Günther *et al*. [20] and evaluate the o_3 probe set. The training set consists of 3 samples for each of the 610 KCs and 1 sample for

each of the 1070 KUCs which is a ratio of 0.58. We extract 128-dimensional features using ResNet-50 [22], pre-trained on MS-Celeb-1M [21], and finetuned on VGGFace2 [9].

CASIA-WebFace (C-WF) [62]. This is a face recognition dataset with imbalanced distributions. We split it into 6345 KCs, 2115 KUCs, and 2115 UUCs. Unless otherwise stated, all experiments are repeated 3 times with varying KC and KUC splits and averaged results are reported. On average, the least represented KC has 13 samples and the largest 627 samples. The KUC to KC sample ratio is 0.33. Feature extraction uses the same model as for LFW but with a feature dimension of 2048. We also employ a reduced version (Tiny C-WF) to address classifier scaling issues with the full dataset. Tiny C-WF retains 20 % of KCs, KUCs, and UUCs. The number of samples per class is reduced by 50 %.

3.2. Evaluating open-set recognition models

First, we assess the ability to distinguish knowns from unknowns in a binary problem. This is supported by a Receiver Operating Characteristic (ROC) depicting the true positive rate (TPR) *vs*. the false positive rate (FPR). To obtain a performance measure independent of the threshold δ, we use the Area Under the ROC Curve (AUC-ROC).

Second, we employ the Open-Set Classification Rate (OSCR) [13], which measures the correct classification rate (CCR) and FPR. Let $g(\boldsymbol{x}, \delta; \mathcal{T})$ be a decision function as formalized in Eq. (1). The test set comprises three types of classes, denoted as $\mathcal{E} = \mathcal{E}_K \cup \mathcal{E}_u$ with $\mathcal{E}_u = \mathcal{E}_{KU} \cup \mathcal{E}_{UU}$. Then these measures are expressed as follows:

$$\mathrm{CCR}(\delta) = |\mathcal{E}_K|^{-1} |\{\boldsymbol{x} | (\boldsymbol{x}, y) \in \mathcal{E}_K \wedge g(\boldsymbol{x}, \delta; \mathcal{T}) = y\}| \,, \quad (2)$$

$$\mathrm{FPR}(\delta) = |\mathcal{E}_u|^{-1} |\{\boldsymbol{x} | \boldsymbol{x} \in \mathcal{E}_u \wedge g(\boldsymbol{x}, \delta; \mathcal{T}) \neq u\}| \,. \quad (3)$$

For most open-set applications, *e. g*. face recognition, a high CCR at low FPR is preferable [20]. Therefore, we report the CCRs in the open-set relevant FPR range of 0 to 10 %.

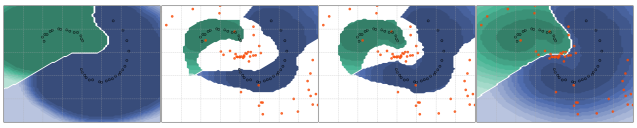
3.3. Biased *vs*. unbiased evaluation

The biased protocol in Section 3.1 with genuine KUCs in the training set simplifies the detection of KUCs in the test set, resulting in significant improvements in OSR measures.

To ensure unbiased evaluation, we compute the metrics solely on a test subset comprising KCs and UUCs, excluding KUCs. Thus, $\mathcal{E}_u$ in Eq. (3) is substituted by $\mathcal{E}_{UU}$. *Biased* evaluation considers metrics for both KUCs and UUCs, while *unbiased* evaluation focuses solely on UUCs. The biased assessment tends to paint an overly optimistic picture for applications that primarily deal with UUCs.

4. How to learn models with known unknowns?

This section outlines three open-set training strategies and their deployment to six classifiers of four categories.



(a) EVM (b) + SPL (c) + MPL (d) + KvR

Figure 2. EVM decision boundaries of training strategies exploiting known unknown classes (KUCs). This is a toy dataset with dark-edged points for 2 known classes (KCs) and bright orange points for KUCs. Colored areas belong to the related class with confidence visualized via opacity, *i. e*., white is *zero* confidence.

4.1. Open-set training strategies

Incorporating KUCs into the training process of classifiers can be achieved using various strategies. Figure 2 illustrates examples for these approaches, including the EVM.

Single Pseudo Label (SPL). Samples without labels are assigned the *pseudo* label u and treated as a single class. This method aligns with the $K+1$ training strategy in previous works [16, 40, 52] with function h, but with $\mathcal{C}_K$ substituted by $\mathcal{C}$. This also implies $f(\boldsymbol{x}; \mathcal{T}) = \boldsymbol{z}$ with $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{C}|}$.

The model can now predict label u directly independent of the threshold δ. This results in steep decision boundaries as demonstrated in Fig. 2b. While this approach is universally applicable, the complex distribution of the large background class presents a challenge. Depending on the underlying OSR model, finding suitable representations for the background class formed by the pseudo label is difficult.

Multi Pseudo Label (MPL). This strategy assigns an individual pseudo label to each KUC sample. The number of classes in $\mathcal{C}$ increases by $|\mathcal{T}_u|$, the amount of KUC samples in $\mathcal{T}$. The remaining strategy follows the SPL, but now each former KUC contains only one sample. If one of the newly declared classes is predicted, it must be mapped to label u.

Similar to SPL, this approach enables direct predictions of unknown labels, resulting in sharp decision boundaries, *cf*. Fig. 2c. We hypothesize that multiple pseudo labels stabilize OSR learning, especially when each KUC sample belongs to a distinct category with high inter-category distances in feature space. A notable drawback is the significant increase in classes, making it impracticable for One *vs*. Rest (OvR) models. Additionally, MPL may induce label noise, leading to inconsistent decision boundaries.

Known *vs*. Rest (KvR). This mode is akin to the common OvR multiclass strategy, often used for binary classifiers like SVMs to handle multiclass problems. In KvR, each KUC never acts as a positive class, avoiding the need for its own model representation. Instead, it serves as negative in the rest-class of other binary models. This approach resolves the issue of complex background class representation. Other classes can still adjust their decision boundaries

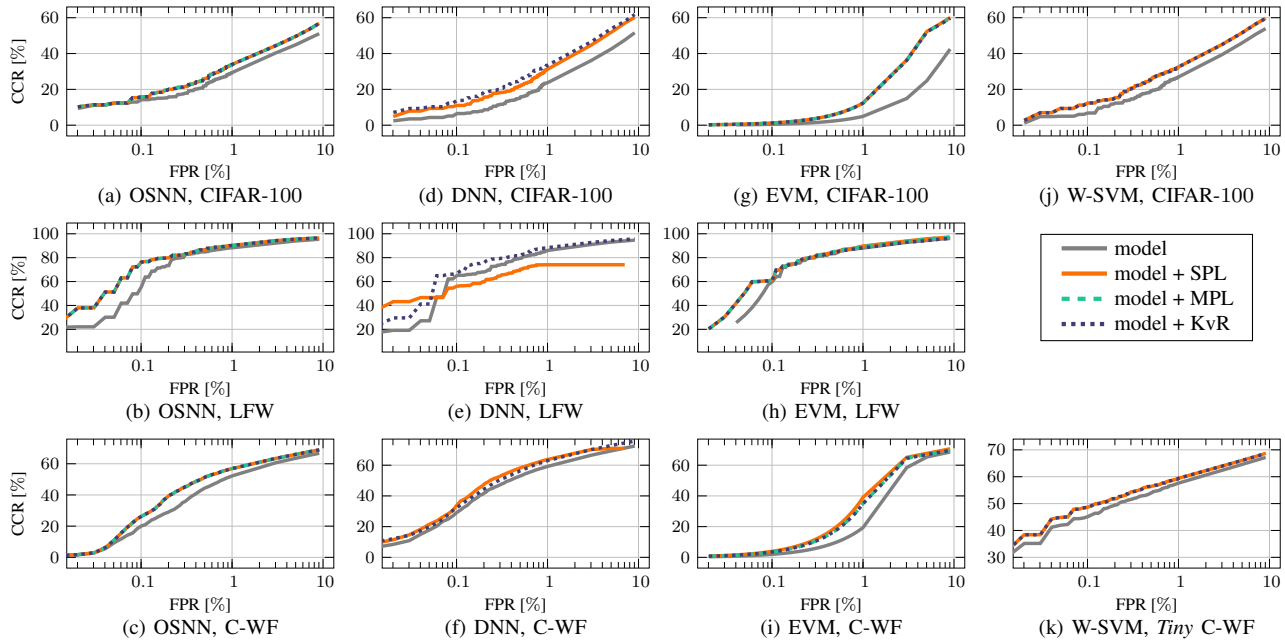


Figure 3. Results of the biased evaluation of 4 models (column-wise) exploiting genuine KUCs with the training strategies and the baseline on 3 datasets (row-wise). The models comprise OSNN in (a) – (c), DNN in (d) – (f), EVM in (g) – (i), and W-SVM in (j) and (k). Results in (k) are for *Tiny* C-WF. Shown is the biased Open-Set Classification Rate (OSCR) in the open-set relevant FPR range up to 10 %.

to the negative samples, observed in Fig. 2d where KCs' boundaries are tightened by KUCs. Implementing KvR can vary, but we intend for the background data not to be explicitly modeled as a pseudo-class. It serves as a tool to determine better decision boundaries for KCs concerning the open space. Experiments in Section 4.3 show that this approach enhances unknown recognition.

4.2. Deployed strategies for open-set models

In this section, we briefly outline how to deploy the strategies to 6 OSR models representing 4 distinct categories. More details are provided in the supplements.

The Open-Set Nearest Neighbor (OSNN) [29] exploits the ratio between the two nearest samples from distinct classes as confidence value. SPL includes all KUCs in distance ratio computation and class prediction. MPL treats each KUC as a separate class, impacting the reject option only. For KvR, only KCs are used for label prediction, while both KCs and KUCs determine distance ratios. The strategies primarily vary in the confidence computation, leading to nearly identical results.

Given the vast possibilities of training Deep Neural Networks (DNNs) with KUCs by tailoring losses, we opt for the classical cross-entropy loss. For SPL, we expand the number of output units by one class. MPL does not scale to large KUC sets. Entropic open-set loss [13] is adopted for KvR.

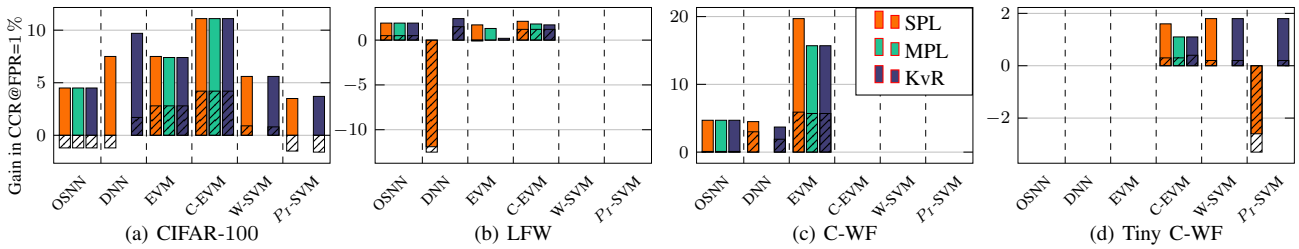
The Extreme Value Machine (EVM) [48], C-EVM [23], and both Support Vector Machine (SVM) variants, Weibull SVM (W-SVM) [51] and Probability of Inclusion SVM (P_I-SVM) [27], implement an OvR scheme. For each *one*-class the other *rest*-classes serve as counterpart. For these models SPL and MPL differ only in the number of pseudo-classes. The SVMs omit MPL due to its limited scalability to large number of classes and their requirement of at least three samples per class. The KvR strategy can be deployed by not representing the KUCs as a positive one-class. Instead, they are always considered part of the rest-class.

4.3. Performance of training strategies

We benchmark the training strategies using the protocols in Section 3.1. Hyperparameters are determined through a grid search based on 5-fold stratified cross-validation. For LFW, standard 3-fold cross-validation is applied. The grid search is conducted for the baseline models and the parameters are reused for SPL, MPL, and KvR. This approach may favour the baseline, where no KUCs are used.

Biased evaluation. In Fig. 3, we report average OSCR measures for the three datasets. Results for the C-EVM and P_I-SVM are reported in the supplement. KUC exploitation improves OSR performance compared to baselines without KUCs. The exact behavior is model-dependent.

For OSNN, *cf*. Figs. 3a – 3c, we observe consistent improvements by SPL, MPL, and KvR. The highest gain of about 21 % is observed in Fig. 3b for FPRs below 0.1 %. This result is crucial for safety-relevant applications that require low FPRs. For C-WF, we see a gain in the FPR range of 0.1 to 1 %, and for CIFAR-100 from 0.1 % upwards. The



(a) CIFAR-100 (b) LFW (c) C-WF (d) Tiny C-WF

Figure 4. Biased *vs*. unbiased OSR performance across all models, strategies, and datasets. Given is the gain in the CCR@FPR=1 % of the strategies using genuine KUCs over the respective baseline model trained without KUCs. The colored bars show the gain in the biased evaluation, *i. e*. the performance over KCs, KUCs, and UUCs. Shaded areas show the gain in the unbiased evaluation with excluded KUCs.

exploitation of KUCs does not negatively impact the CCR.

For DNN, *cf*. Figs. 3d – 3f, KvR consistently outperforms the baseline. SPL performs similarly to KvR for CIFAR-100 and C-WF but is outperformed by KvR and the baseline on LFW, except for FPRs below 0.1 %. We explain this behaviour by the few samples per KCs and KUC in case of LFW, causing two problems: 1) Each KUC is represented by one sample, making it challenging to model a pseudo-class. 2) The pseudo-class is large compared to the KCs. After examining SPL, we found that the DNN mainly predicts unknowns, a common issue of DNNs with unbalanced classes [8]. This is not the case for C-WF and CIFAR-100, where all classes have sufficient samples.

For the EVM in Figs. 3g – 3i, all training strategies outperform the baseline. For CIFAR-100, we observe a gain of 17.6 % in the CCR@FPR=10 %, for LFW a gain of 30.6 % at an FPR of 0.05 %, and for C-WF 20 % at an FPR of 1 %.

The W-SVM results for CIFAR-100 and Tiny C-WF are in Figs. 3j and 3k. As indicated in Section 4.2, we do not apply the MPL strategy and also do not evaluate LFW due to the model's limitations. By exploiting KUCs, OSR performance in the selected open-set range is improved by up to 6 % and 4 % for CIFAR-100 and Tiny C-WF, respectively.

Unbiased evaluation. The question remains whether KUCs also improve the detection of UUCs. In Fig. 4, we compare the biased and unbiased evaluation using the CCR@FPR=1 % gain in relation to the baseline. The performances expressed by the unbiased evaluation fall below the biased measures. This indicates that UUCs are more difficult to detect and the biased evaluation draws a too optimistic picture. We found that this discrepancy is model-dependent. The OSNN exhibits a loss for CIFAR-100 in the unbiased measure and only marginal gains for LFW and C-WF. The DNN with KvR outperforms the baseline in both evaluations. This supports the finding that the entropic open-set loss is more suitable for OSR than the simple K+1 (SPL) strategy [13]. The EVMs improves over the baselines in both biased and unbiased cases, except in for LFW. Unlike the OSNN, the EVMs take advantage of a tail of KC and KUC samples to refine the decision boundary of KCs. In contrast to the W-SVM, the P_I-SVM does not always benefit from KUCs in the unbiased case.

5. How to generate known unknowns?

We showed that genuine KUCs can improve OSR. However, in certain domains, genuine KUC data may be unavailable. This raises the question of how to synthesize them. In contrast to GANs that are not easily usable as standalone OOD generator for arbitrary OSR models, *cf*. Section 2.2, our proposed framework answers this question using the tools at hand, namely the data and the feature extractor.

5.1. Mixup – known unknowns straight off the shelf

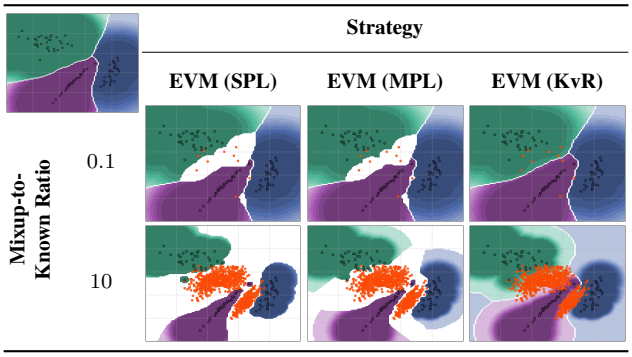
Mixup [66] is a lightweight augmentation technique for DNNs that generates new data by convex combinations of samples. Jian *et al*. [28] show that standard mixup increases open-space risk by smoothing decision boundaries between classes. However, we further constrain mixup to use samples from distinct classes and assign to the mixed sample $\tilde{\boldsymbol{x}}$ the unknown label u, leading to steep decision boundaries. Here, $(\boldsymbol{x}_i, y_i)$ and $(\boldsymbol{x}_j, y_j)$ are randomly drawn from $\mathcal{T}_K$:

$$\tilde{\boldsymbol{x}} = \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_j \quad \text{, with } \; y_i \neq y_j \; \text{ and } \; \tilde{y} = u \;\; . \tag{4}$$

We use *manifold mixup* [57], as mixing in feature space offers several advantages: 1) It is more efficient, as the combinations no longer need to be computed in the high-dimensional image space. 2) It is independent of peculiarities of the feature extractor, eliminating the need to pass the mixed image through that extractor.

We use CIFAR-100 and replace genuine KUCs with mixups of KCs. Mixing factor λ in Eq. (4) is set to $\lambda \sim \mathrm{Beta}(2,2)$ with $\lambda \in [0.4, 0.6]$, ensuring that most mixups lie between two classes. While training with synthetic and testing with genuine KUCs is not biased, we present the metrics calculated on the test set with excluded KUCs for a fair comparison with models using genuine KUCs, and we continue to refer to this as *unbiased*. We also assess OSR performance by the number of generated mixup samples, represented by the *mixup-to-known ratio*, indicating how many mixups are generated per one known sample.

Table 1. EVM class boundaries with the baseline in the top left and applied strategies. This toy dataset contains dark-edged dots from 3 KCs and orange dots as mixups. Colored areas display class assignment, with opacity indicating confidence, where white is zero confidence or, conversely, high confidence for open space.



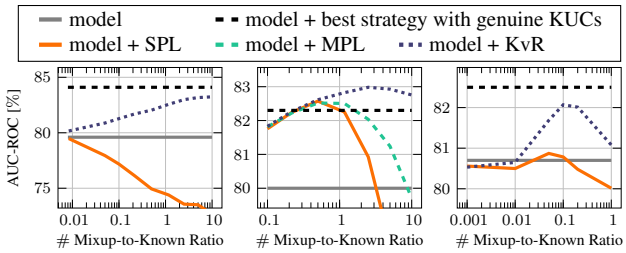
5.2. Augmenting models by manifold mixup

We exemplary investigate manifold mixup for the EVM in Tab. 1. With more mixups, the gaps between classes are filled and the open space is extended. The strategies have varying degrees of invasiveness, with SPL favoring open space the most, followed by MPL and then KvR. KvR requires finding a suitable threshold to form open space.

The AUC-ROC results for naïvely generated off-the-shelf mixups are shown in Fig. 5. Other metrics like the OSCR are anticipated here and can be found in the supplement for a more detailed analysis, this includes also the results for the C-EVM and P_I-SVM.

Similar to the findings in the first unbiased experiment in Fig. 4, the DNN (SPL) in this experiment also does not benefit from the additional background data, *cf*. Fig. 5a. The AUC-ROC and CCR@FPR=10 % of KvR show a steady increase. Interestingly, this does not apply to the entire FPR range, as the CCR@FPR=1 % temporarily deteriorates. This suggests that mixups may offer advantages at medium FPRs while providing no benefit at low FPRs, or vice versa. This makes it difficult to make a universal statement, such as mixup always improves OSR performance.

The EVM in Fig. 5b shows remarkable benefits from all strategies and mixup. The AUC-ROC corresponds perfectly with Tab. 1. SPL constricts the space for the KCs more than MPL, resulting in higher recognition rates of unknowns at the expense of knowns. MPL and KvR loosen this constriction. Not only the baseline, but also EVM (SPL) trained with genuine KUCs is outperformed. This is an intriguing observation, showing that mixup improves UUCs recognition to the same extent or even more than genuine dataset-related data, which is often expensive to obtain in real-world scenarios [59]. Although SPL is slightly better in the first experiment, *cf*. Fig. 4, we recommend KvR as preferred strategy due to its consistently stable results across all FPRs.



(a) DNN, best = KvR (b) EVM, best = SPL (c) W-SVM, best = KvR

Figure 5. Unbiased CIFAR-100 results of models exploiting mixups. Shown are the 3 strategies with the AUC-ROC over the number of mixup samples. The baseline model without KUCs (—) and the best strategy of each model trained with genuine KUCs (- -) from the first unbiased experiment, *cf*. Fig. 4, serve as reference. Best strategies are indicated in the respective subtitle.

The result of the W-SVM is in Fig. 5c. Training with KvR yields a comparable AUC-ROC gain to genuine KUCs. With a mixup-to-known ratio of 0.1, it exceeds the baseline by about 1.2 %.

Improving OSR with mixup is not universally effective. The OSNN and P_I-SVM, *cf*. supplement, decline in performance when mixup is applied. A nearest neighbor based approach is particularly prone to unfavorable sample placement, which is exacerbated by naïve mixups. We address this occupation problem in the next section.

5.3. Solving the occupation problem

Zhou *et al*. [67] mention the issue of already-occupied space between two KCs by a third. Since mixups are labeled as unknown, this can degrade the adjacent KCs. As augmentation technique or in feature learning, the occupation problem is not approached due to two reasons: 1) Mixups are generated in mini-batches and checking on their location in the feature space in relation to *all* other samples is infeasible. 2) The occupation might be favorable as mixups push away the in-between class, increasing the space between all affected classes. However, in case of pre-trained features, this can indeed impact OSR performance. We hypothesize that it is the main reason for the OSNN failure case.

We propose an effective constraint for an on-the-fly mixup generation. Let $\bar{\boldsymbol{x}}_y$ be a KC's centroid, calculated as mean of all samples within a class. Using Euclidean distance $d(\cdot,\cdot)$, we determine the mean distance $\bar{d}$ among all class centroids. A mixup sample is kept if its distance to all class centroids exceeds the scaled $\bar{d}$. Scale $\alpha = 0$ represents unconstrained generation and $\alpha > 0$ promotes filtering:

$$d(\tilde{\boldsymbol{x}}, \bar{\boldsymbol{x}}_y) > \alpha \cdot \bar{d} \quad \forall y \in \mathcal{C}_K \ . \tag{5}$$

We conduct the CIFAR-100 experiment again, varying $\alpha \in \{0, 0.6, 0.8, 1\}$. Results are presented in Fig. 6 and we recommend to view the supplement for additional details.

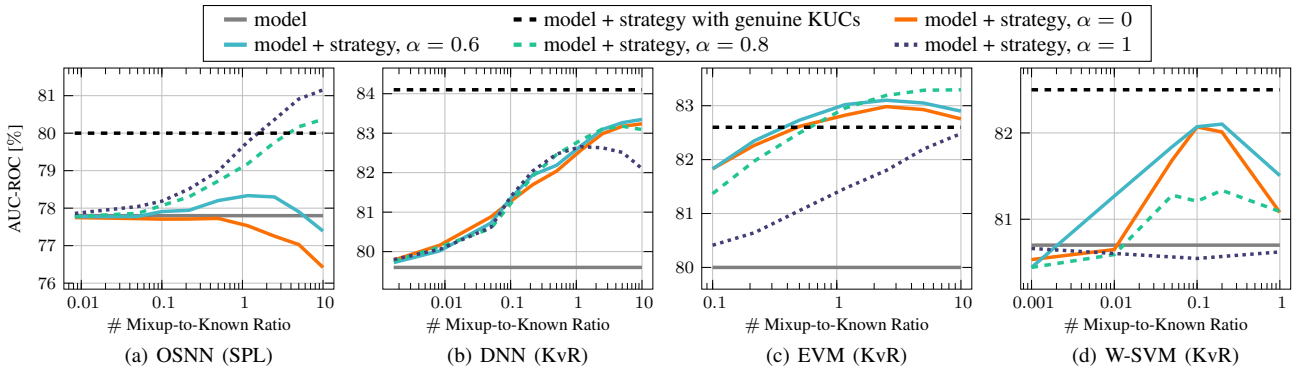


Figure 6. Unbiased results of the models exploiting constrained mixups on CIFAR-100. Shown is one strategy for each model with the AUC-ROC over the number of mixup samples. The baseline model (—) and the model exploiting genuine KUCs (- -) serve as reference.

The OSNN in Fig. 6a benefits from stronger constraints ($\alpha \geq 0.8$), resulting in consistent AUC-ROC improvements of up to $3.4\,\%$ compared to the baseline as the number of mixups increases. The CCR@FPR=$1\,\%$ is erratic, but with FPR=$10\,\%$, it improves by almost $2\,\%$. We note three observations: 1) OSNN is prone to naïve mixups. 2) Highly constrained mixups improve OSR, but not across the entire FPR range. 3) Performance deterioration is unlikely and can be mitigated by monitoring a held-out set.

The DNN (KvR) in Fig. 6b has little benefit with stronger restrictions, for $\alpha = 1$ it even deteriorates after a mixup-to-known ratio of 1. Mixups lead to an improvement of up to $4\,\%$, especially for FPRs $>1\,\%$. We conclude: 1) Naïve mixups enhance OSR for medium and larger FPRs. 2) DNNs usually benefit from diverse data, constraining mixups might be counterproductive.

The EVM (KvR) in Fig. 6c shows that no or lower constraints ($\alpha < 0.8$) increase more with the number of mixup samples in the CCR than the more restricted ones. The magnitude of the constraints affects different areas of the FPR. Our conclusions are: 1) Due to the variable behavior of different constraints, the EVM can be optimized for a desired FPR. 2) Finding the optimal configuration is complex and may require a hyperparameter search.

While the AUC-ROC of the W-SVM in Fig. 6d appears promising, the OSCR is irregular with a downward trend. This behaviour is underlined by the toy example in the supplement, where different mixup-to-known ratios lead to erratic changes at low confidences.

6. Discussion and Conclusion

This paper presents LORD to explicitly model open space by exploiting KUCs for open-set learning and to boost OSRs models. Our key findings are as follows:

Known unknowns improve OSR learning. Exploiting KUCs to train OSR models improves over baselines ignoring such samples. Considering KUCs improves the detection of KUCs and UUCs. The performance for UUCs is usually lower. Practitioners should carefully select the evaluation strategy and only benchmark with KUCs if this reflects the conditions in the addressed use case.

The training strategy matters. SPL can be considered as a baseline due to its simplicity. When used with synthetic KUCs, SPL is outperformed by competing methods. MPL shows comparable results but is not tractable for every model, treating each KUC sample as a separate class. KvR outperforms the other strategies in most benchmarks and requires fewer parameters to learn.

The OSR evaluation measure matters. AUC-ROC measures the ability to distinguish knowns and unknowns, but relying solely on this metric can create false expectations. The relevant FPR range for OSR is usually $<10\,\%$, occupying a small part of the AUC and being outweighed by the range $>10\,\%$. AUC-ROC improvements might be due to enhancements at high FPRs, which have limited relevance for OSR. It is crucial to report TPR or CCR at low FPRs. For future work, the OpenAUC [60] could also be a helpful alternative to measure OSR performance.

Mixups can surrogate genuine known unknowns. Mixups serve as a lightweight and effective surrogate for KUCs. DNN, EVM, and OSNN benefit the most from mixups. OSNN requires extra filtering to mitigate the occupation problem. Filtering does not consistently improve all models but increases the CCR in certain FPR ranges. The aim of solving the occupation problem is to maintain a high KC classification. A suitable compromise between unknown detection and known recognition has to be found.

LORD is extendable to open-world learning. Future work refines the generation of known unknowns. The constraints to address the occupation problem shall be enhanced. The highly efficient mixup synthesis proposed here is suitable for open-world learning and future experiments should involve open-world capable classifiers.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access*, 6:14410–14430, 2018. 3

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 2

[3] Peter L. Bartlett and Marten H. Wegkamp. Classification with a Reject Option Using a Hinge Loss. *Journal of Machine Learning Research (JMLR)*, 9(8):1823–1840, 2008. 1

[4] Abhijit Bendale and Terrance E. Boult. Towards Open World Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1893–1902, 2015. 1

[5] Abhijit Bendale and Terrance E. Boult. Towards Open Set Deep Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1563–1572, 2016. 1

[6] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Network. In *International Conference on Machine Learning (ICML)*, pages 1613–1622. PMLR, 2015. 1

[7] Terrance E. Boult, Nicolas M. Windesheim, Steven Zhou, Christopher Pereyda, and Lawrence B. Holder. Weibull-Open-World (WOW) Multi-Type Novelty Detection in CartPole3D. *Algorithms*, 15(10):381, 2022. 1

[8] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. *Neural Networks*, 106:249–259, 2018. 6

[9] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A Dataset for Recognising Faces Across Pose and Age. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, Xi'an, China, May 2018. 4

[10] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning Augmentation Strategies from Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123, 2019. 3

[11] Amirabbas Davari, Erchan Aptoula, Berrin Yanikoglu, Andreas Maier, and Christian Riess. GMM-Based Synthetic Samples for Classification of Hyperspectral Images with Limited Training Data. *IEEE Geoscience and Remote Sensing Letters (GRSL)*, 15(6):942–946, 2018. 3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009. 3

[13] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing Network Agnostophobia. *Advances in Neural Information Processing Systems (NIPS)*, 31, 2018. 2, 3, 4, 5, 6, 12

[14] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[15] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust Physical-World Attacks on dDep Learning Visual Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1625–1634, 2018. 3

[16] Zongyuan Ge, Sergey Demyanov, and Rahil Garnavi. Generative OpenMax for Multi-Class Open Set Classification. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 42.1–42.12. BMVA Press, 2017. 2, 3, 4

[17] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent Advances in Open Set Recognition: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 43(10):3614–3631, 2020. 1

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. *Advances in Neural Information Processing Systems (NIPS)*, 27, 2014. 2, 3

[19] Yves Grandvalet, Alain Rakotomamonjy, Joseph Keshet, and Stéphane Canu. Support Vector Machines with a Reject Option. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS)*, 2008. 1

[20] Manuel Günther, Steve Cruz, Ethan M. Rudd, and Terrance E. Boult. Toward Open-Set Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 71–80, 2017. 1, 2, 3, 4

[21] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *European Conference on Computer Vision (ECCV)*, pages 87–102. Springer, 2016. 4

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[23] James Henrydoss, Steve Cruz, Chunchun Li, Manuel Günther, and Terrance E. Boult. Enhancing Open-Set Recognition Using Clustering-Based Extreme Value Machine (C-EVM). In *International Conference on Big Data (BigData)*, pages 441–448. IEEE, 2020. 1, 5, 12

[24] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. SimPLE: Similar Pseudo Label Exploitation for Semi-Supervised Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15099–15108, 2021. 2

[25] Gary B. Huang and Erik Learned-Miller. Labeled Faces in the Wild: Updates and New Reporting Procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, MA, USA, 2014. 3

[26] Gary B. Huang, Marwan Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environ-

ments. Technical Report UM-CS-07-49, University of Massachusetts, Amherst, MA, USA, 2007. 3

[27] Lalit P. Jain, Walter J. Scheirer, and Terrance E. Boult. Multi-Class Open Set Recognition Using Probability of Inclusion. In *European Conference on Computer Vision (ECCV)*, pages 393–409. Springer, 2014. 1, 5, 12

[28] Guosong Jiang, Pengfei Zhu, Yu Wang, and Qinghua Hu. OpenMix+: Revisiting Data Augmentation for Open Set Recognition. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2023. 6

[29] Pedro R. Mendes Júnior, Roberto M. De Souza, Rafael de O. Werneck, Bernardo V. Stein, Daniel V. Pazinato, Waldir R. de Almeida, Otávio A. B. Penatti, Ricardo da S. Torres, and Anderson Rocha. Nearest Neighbors Distance Ratio Open-Set Classifier. *Springer Machine Learning (ML)*, 106(3):359–386, 2017. 1, 5, 12

[30] Tobias Koch, Felix Liebezeit, Christian Riess, Vincent Christlein, and Thomas Köhler. Exploring the Open World Using Incremental Extreme Value Machines. In *International Conference on Pattern Recognition (ICPR)*, pages 2792–2799, 2022. 1

[31] Shu Kong and Deva Ramanan. OpenGAN: Open-Set Recognition via Open Data Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 813–822, 2021. 2, 3

[32] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009. 3

[33] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast AutoAugment. *Advances in Neural Information Processing Systems (NIPS)*, 32, 2019. 3

[34] Benedikt Lorch, Anatol Maier, and Christian Riess. Reliable JPEG Forensics via Model Uncertainty. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2020. 1

[35] Benedikt Lorch, Franziska Schirrmacher, Anatol Maier, and Christian Riess. Reliable Camera Model Identification Using Sparse Gaussian Processes. *IEEE Signal Processing Letters (SPL)*, 28:912–916, 2021. 1

[36] Michael McCloskey and Neal J Cohen. Catastrophic Interference in Connectionist Nnetworks: The Sequential Learning Problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier, 1989. 3

[37] Stefano Melacci and Mikhail Belkin. Laplacian Support Vector Machines Trained in the Primal. *Journal of Machine Learning Research (JMLR)*, 12(3):1149–1184, 2011. 2

[38] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-Based Image Classification: Generalizing to New Classes at Near-Zero Cost. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(11):2624–2637, 2013. 1

[39] Dimity Miller, Niko Sunderhauf, Michael Milford, and Feras Dayoub. Class Anchor Clustering: A Loss for Distance-Based Open Set Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3570–3578, 2021. 1

[40] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open Set Learning with Counterfactual Images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 613–628, 2018. 2, 3, 4

[41] Poojan Oza and Vishal M. Patel. C2AE: Class Conditioned Auto-Encoder for Open-Set Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2307–2316, 2019. 1

[42] Pramuditha Perera, Vlad I. Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M. Patel. Generative-Discriminative Feature Representations for Open-Set Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11814–11823, 2020. 1

[43] Pramuditha Perera and Vishal M. Patel. Deep Transfer Learning for Multiple Class Novelty Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11544–11552, 2019. 2, 3

[44] Pramuditha Perera and Vishal M. Patel. Geometric Transformation-Based Network Ensemble for Open-Set Recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 3

[45] Luis Perez and Jason Wang. The Effectiveness of Data Augmentation in Image Classification Using Deep Learning. *preprint arXiv:1712.04621*, 2017. 3

[46] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. iCaRL: Incremental Classifier and Representation Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2001–2010, 2017. 3

[47] Marko Ristin, Matthieu Guillaumin, Juergen Gall, and Luc Van Gool. Incremental Learning of NCM Forests for Large-Scale Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3654–3661, 2014. 1

[48] Ethan M. Rudd, Lalit P. Jain, Walter J. Scheirer, and Terrance E. Boult. The Extreme Value Machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(3):762–768, 2017. 1, 5, 12, 13

[49] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A Unified Survey on Anomaly, Novelty, Open-Set, and Out-of-Distribution Detection: Solutions and Future Challenges. *preprint arXiv:2110.14051*, 2021. 1

[50] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(7):1757–1772, 2012. 1, 3

[51] Walter J. Scheirer, Lalit P. Jain, and Terrance E. Boult. Probability Models for Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(11):2317–2324, 2014. 1, 5, 12

[52] Patrick Schlachter, Yiwen Liao, and Bin Yang. Open-Set Recognition Using Intra-Class Splitting. In *27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019. 2, 3, 4

[53] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from Simulated and Unsupervised Images through Adversarial Training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2107–2116, 2017. 3

[54] Lei Shu, Hu Xu, and Bing Liu. DOC: Deep Open Classification of Text Documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2911–2916, 2017. 1

[55] Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning (ICML)*, pages 6105–6114. PMLR, 2019. 3

[56] David MJ Tax and Robert PW Duin. Growing a Multi-Class Classifier with a Reject Option. *Pattern Recognition Letters*, 29(10):1565–1570, 2008. 1

[57] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold Mixup: Better Representations by Interpolating Hidden States. In *International Conference on Machine Learning (ICML)*, pages 6438–6447. PMLR, 2019. 3, 6

[58] Edoardo Vignotto and Sebastian Engelke. Extreme Value Theory for Open Set Classification – GPD and GEV Classifiers. *preprint arXiv:1808.09902*, 2018. 1

[59] Dong Wang, Yuan Zhang, Kexin Zhang, and Liwei Wang. FocalMix: Semi-Supervised Learning for 3D Medical Image Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3951–3960, 2020. 2, 7

[60] Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. OpenAUC: Towards AUC-Oriented Open-Set Recognition. *Advances in Neural Information Processing Systems (NIPS)*, 35:25033–25045, 2022. 8

[61] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized Out-of-Distribution Detection: A Survey. *preprint arXiv:2110.11334*, 2021. 1

[62] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning Face Representation from Scratch. *preprint arXiv:1411.7923v1*, 2014. 4

[63] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-Reconstruction Learning for Open-Set Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4016–4025, 2019. 1

[64] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019. 3

[65] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. *Advances in Neural Information Processing Systems (NIPS)*, 34:18408–18419, 2021. 2

[66] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 3, 6

[67] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning Placeholders for Open-Set Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2021. 3, 7

A. Appendix

This supplementary material is organized as follows:

- Appendix A.1 provides some further details on the implementation of the strategies in the models.
- Appendix A.2 presents the results from the Clustering-based EVM (C-EVM) and P_I-SVM [27], exploiting genuine KUCs with the learning strategies.
- Appendix A.3 shows visualizations of the decision boundaries on a toy dataset for all models and various mixup-to-known ratios. It also includes additional OSR measures for the models in the main work, the C-EVM, and the P_I-SVM, using mixups as KUC surrogates.
- Appendix A.4 contains additional metrics for the models in the main work, the C-EVM, and the P_I-SVM, with constrained mixups.

A.1. Deployed strategies for open-set models – additional details

In this section, we outline how to deploy the strategies to 6 different OSR models from the following 4 categories.

Open-Set Nearest Neighbor (OSNN). The OSNN [29] exploits the ratio between the two nearest samples from distinct classes as confidence value. Let d_i and d_j be Euclidean distances between a query sample $\boldsymbol{x}$ and its two closest training samples, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where $y_i \neq y_j$. The ratio is computed as $r = {}^{d_i}/_{d_j}$ with $d_i < d_j$. If $r \leq \delta$, query $\boldsymbol{x}$ is labeled as y_i, otherwise it is labeled as u.

SPL includes all KUCs in distance ratio computation and class prediction. MPL treats each KUC as a separate class, impacting the reject option only. For KvR, the distance ratio is set to the maximum $r = 1$ if the nearest sample $\boldsymbol{x}_i$ belongs to a KUC. Thus, only KCs are used for label prediction, while both KCs and KUCs determine distance ratios. The strategies primarily vary in the confidence computation, leading to nearly identical results.

Deep Neural Network (DNN). Given the vast possibilities of using KUCs in training DNNs by tailoring losses, we opt for the classical cross-entropy loss. The feature extractors mentioned in the dataset-specific paragraphs of the main manuscript are used. A single fully-connected layer with softmax activation is attached and finetuned.

For SPL, we expand the number of output units by one class. We do not evaluate MPL as it does not scale to large KUCs sets. For KvR, we note that this strategy is equivalent to the entropic open-set loss [13]. The objective of entropic open-set is to predict a uniform distribution of unknowns, while KCs are learned according to a cross-entropy loss.

Extreme Value Machines (EVMs). The EVM [48] estimates a Weibull distribution for each sample, considering distances to the nearest samples from other classes. It implements an OvR scheme as it uses distances to *rest*-class samples for each *one*-class sample. The τ smallest distances termed *tail* are used to estimate Weibull distributions.

SPL and MPL use 1 or $|\mathcal{T}_u|$ pseudo-classes, respectively. Unlike SPL, MPL allows KUCs in the tail of other KUCs, causing regularization among close KUCs. This effect is visible when comparing SPL and MPL in Tab. A.2, where the space between KCs and densely populated KUC areas is prioritized for KCs. KvR uses KUCs only in the tails of KCs, and KUCs never act as one-class themselves, leading to gentle transitions between decision boundaries.

We also explore the C-EVM [23], which employs DBSCAN clustering on a per-class basis before the EVM fitting. Clustering calculates centroids for each cluster, serving as proxies for all samples within the cluster. EVM fitting is exclusively performed on these centroids. This model-agnostic preprocessing technique can be applied to any other method mentioned in this work. One particular aspect of interest is whether clustering can counteract the label noise of the MPL strategy. For SPL and MPL, clustering is applied to the entire KUCs as a unified class. For MPL, the resulting class centroids are treated as independent classes again, replacing redundant KUCs and avoiding label noise. For KvR learning, only the KCs undergo cluster-based reduction while leaving the KUCs unaffected.

Support Vector Machines (SVMs). We deploy the training strategies to two SVM variants. The Weibull SVM (W-SVM) [51] combines a one-class SVM and a binary OvR SVM for each class. Weibull distributions are estimated from both SVMs and probabilities are determined by these Weibull distributions. The Probability of Inclusion SVM (P_I-SVM) [27] predicts unnormalized posterior inclusion probabilities with RBF kernels. Weibull distributions are estimated on samples near decision boundaries.

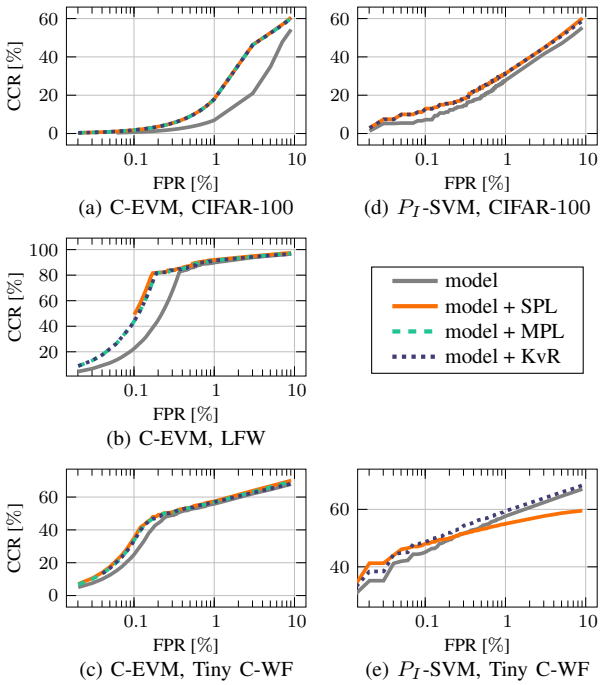
The training strategies SPL and MPL differ only in the number of pseudo-classes. We omit MPL due to its limited scalability to large number of classes. Also, MPL is makes it challenging to serve the SVMs' requirement of at least three samples per class (preferably more). For the same reason, we do not evaluate the SVMs on LFW. The KvR strategy can be deployed to the SVMs by not representing the KUCs as a positive one-class. Instead, they are always considered part of the rest-class during training.

A.2. Performance of training strategies – additional results

This section complements the experiments in which genuine KUCs are exploited by the three learning strategies: 1) SPL, 2) MPL, and 3) KvR.

Figures A.1a – A.1c show the biased OSCR of the C-EVM within the open-set relevant FPR range on CIFAR-100, LFW, and Tiny C-WF. The C-EVM demonstrates similar behavior to the vanilla EVM [48]. All strategies perform comparably well and outperform the baseline. However, it remains inconclusive whether the prior clustering of background data in SPL and MPL prevents potential label noise, leading to improved detection. Two possible conclusions arise: 1) In all three datasets, there is no noise within the genuine KUCs. 2) The C-EVM performs well even without prior clustering of background data, as in KvR.

Figures A.1d and A.1e show the biased results of the P_I-SVM. For CIFAR-100, consistent improvement over the baseline is evident. However, for Tiny C-WF, SPL at FPRs greater than $1\,\%$ results in a degradation of the CCR. The P_I-SVM appears to encounter challenges in modeling all KUCs within a single class, whereas the W-SVM in the main work performs well. This discrepancy might be attributed to the W-SVM's use of a one-class and a binary SVM for each class, while the P_I-SVM deals solely with a binary SVM.



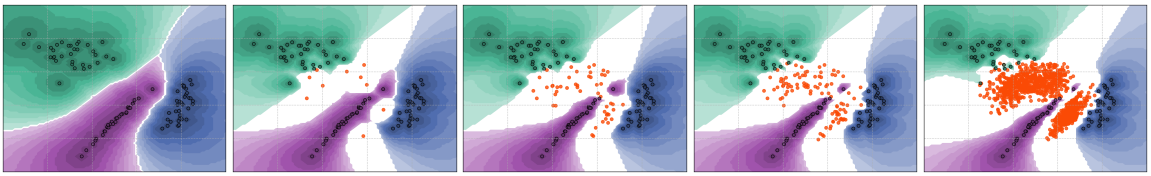
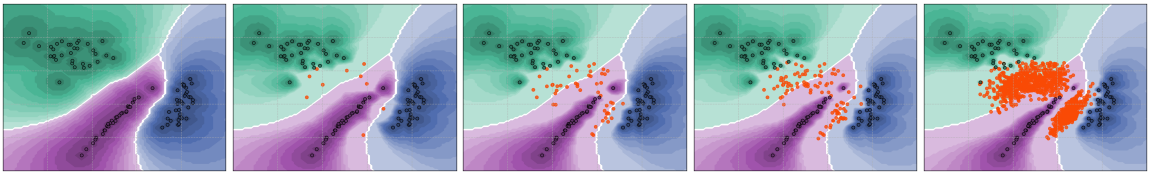
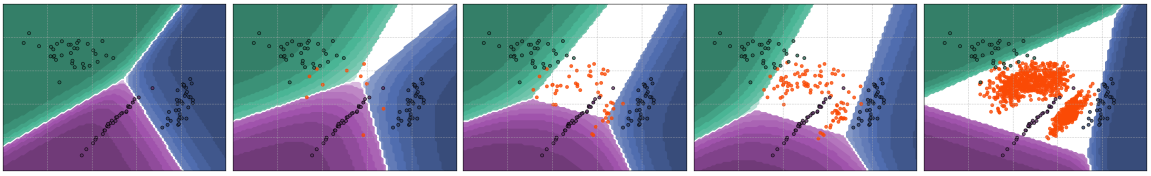
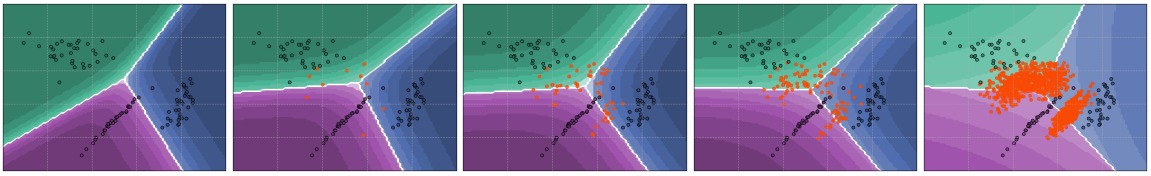
(a) C-EVM, CIFAR-100 (d) P_I-SVM, CIFAR-100

(b) C-EVM, LFW

(c) C-EVM, Tiny C-WF (e) P_I-SVM, Tiny C-WF

Figure A.1. Results of the biased evaluation of 2 models (column-wise) exploiting genuine KUCs with the strategies and the baseline on 3 datasets (row-wise). The models are C-EVM in (a) – (c) and P_I-SVM in (d) and (e). Shown is the biased Open-Set Classification Rate (OSCR) in the open-set relevant FPR range up to $10\,\%$.

Table A.1. OSNN (top) and DNN (bottom) class boundaries with applied strategies and a column-wise increase in mixup samples. This toy dataset contains dark-edged dots from 3 known classes (KCs) and orange dots as mixups. Colored areas display class assignment, with opacity indicating confidence, where white is zero confidence or, conversely, high confidence for open space.

Strategy	**# Mixup-to-Known Ratio**				
	0	0.1	0.5	1	10
OSNN (SPL) & OSNN(MPL)					
OSNN (KvR)					
DNN (SPL)					
DNN (KvR)					

A.3. Augmenting models by manifold mixup – toy examples and additional results

In this section, we present additional decision boundary illustrations using a toy example and various mixup-to-known ratios. The second paragraph contains additional open-set measures for the experiment involving naïve mixup samples as KUCs surrogates.

Toy example visualizations. Tab. A.1 shows the behavior of the OSNN and DNN with the deployed strategies. As observed in the main manuscript, the strategies exhibit minimal differences in OSNN. The model only considers two nearest neighbors, leaving limited scope for variation.

For the DNN, the open space expands as the number of mixups increases, consequently pushing the decision boundaries closer to the known classes. In instances where SPL encounters unfavorably located classes, such as the purple one, the outcome can be very unfavorable. In contrast, the KvR approach is more lenient and consistently offers more flexibility by employing an appropriate threshold.

Tab. A.2 shows both EVM variants with the toy example. As the vanilla EVM has already been discussed in the main manuscript, this section displays the remaining mixup-to-known ratios.

Unlike other methods, the training data for the C-EVM is shown here after the cluster-based reduction. As described in Appendix A.1, SPL and MPL treat KUCs as a single class during clustering, potentially reducing label noise in MPL. In KvR, the background data is not reduced. Notably, we observe a concave-like function when combining oversampling mixups with cluster-based reduction. Oversampling generates larger coherent clusters, leading to a decrease in the number of clusters beyond a certain point. Each cluster is reduced to a centroid, resulting in the observation that more mixups lead to fewer mixups in this type of reduction.

Tab. A.3 displays the toy example alongside the W-SVM and P_I-SVM. The decision boundaries of both SVMs are generally comparable with only slight variations. In the low-confidence range, there are occasional abrupt changes. However, since the nearly transparent areas correspond to very low confidence values, a suitable threshold would usually consider these areas as open space.

Table A.2. EVM (top) and C-EVM (bottom) class boundaries with applied strategies and a column-wise increase in mixup samples. This toy dataset contains dark-edged dots from 3 KCs and orange dots as mixups. Note that for the C-EVM the visible training and mixup dots are the remaining samples *after* the cluster-based reduction. Only in KvR are the mixup samples not reduced. Colored areas display class assignment, with opacity indicating confidence, where white is zero confidence or, conversely, high confidence for open space.

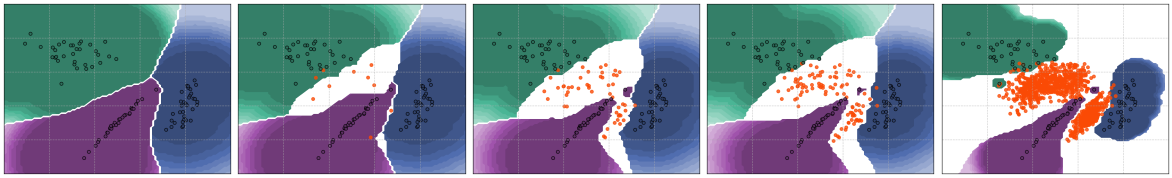
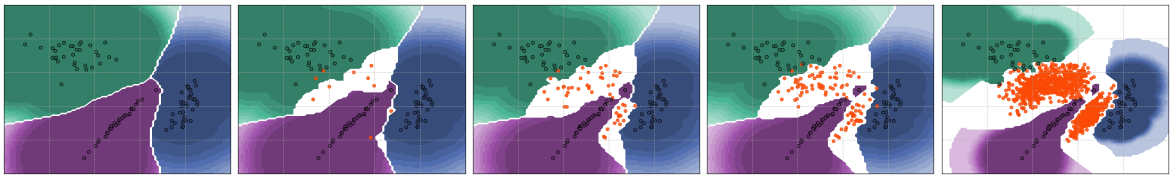
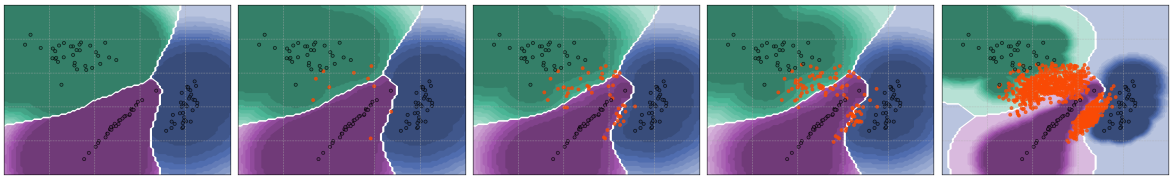
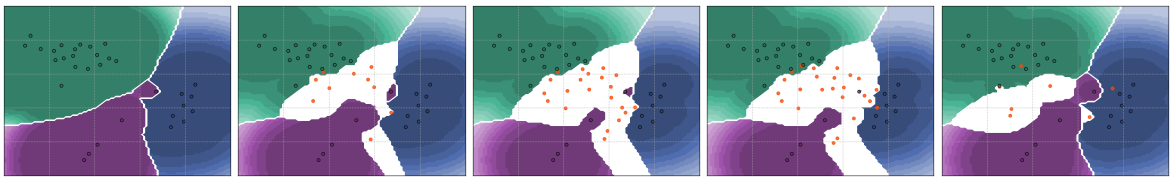
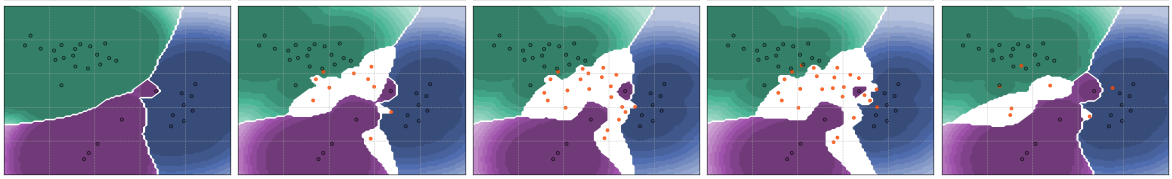
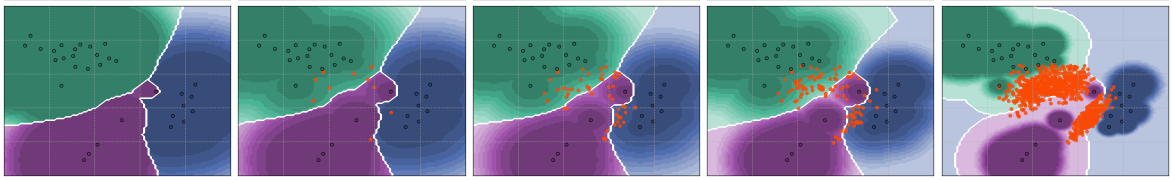
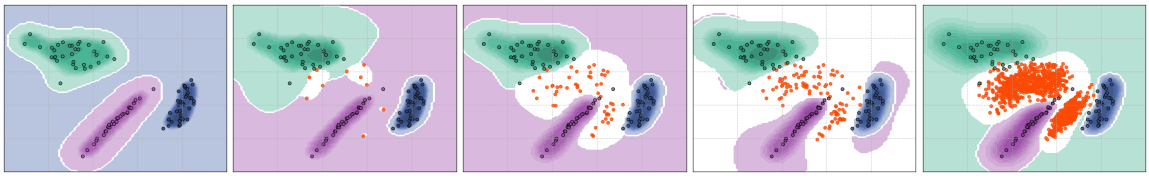
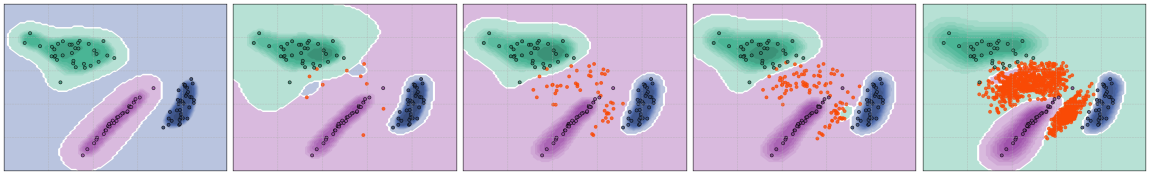
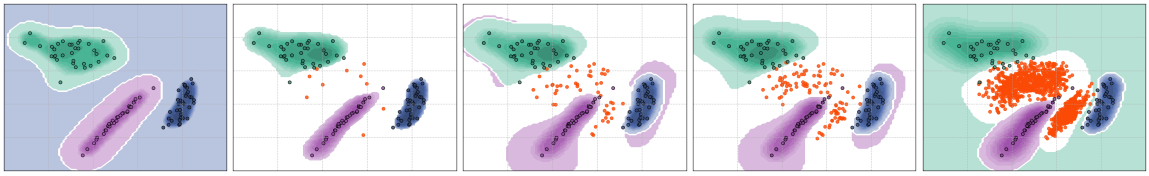
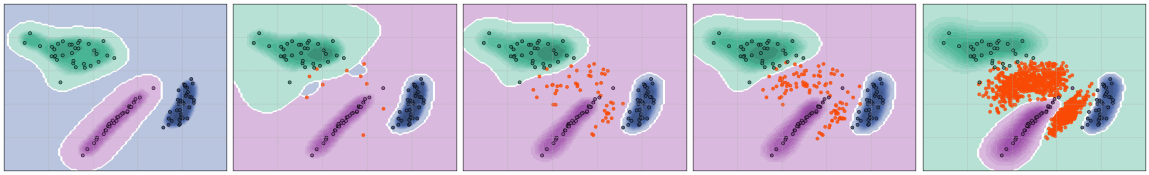
Strategy	**# Mixup-to-Known Ratio**				
	0	0.1	0.5	1	10
EVM (SPL)					
EVM (MPL)					
EVM (KvR)					
C-EVM (SPL)					
C-EVM (MPL)					
C-EVM (KvR)					

Table A.3. W-SVM (top) and P_I-SVM (bottom) class boundaries with applied strategies and a column-wise increase in mixup samples. This toy dataset contains dark-edged dots from 3 known classes (KCs) and orange dots as mixups. Colored areas display class assignment, with opacity indicating confidence, where white is zero confidence or, conversely, high confidence for open space.

Strategy	**# Mixup-to-Known Ratio**				
	0	0.1	0.5	1	10
W-SVM (SPL)					
W-SVM (KvR)					
P_I-SVM (SPL)					
P_I-SVM (KvR)					

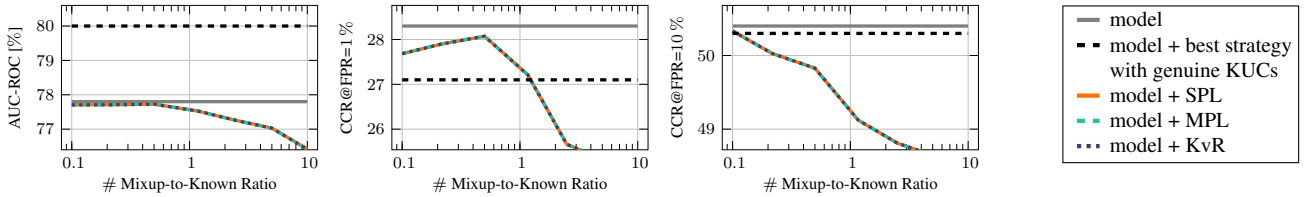
Additional results. This paragraph completes the experiments involving the replacement of genuine KUCs with mixup samples. Figure A.2 combines additional open-set measures for the methods discussed in the main manuscript, along with the remaining results for the C-EVM and P_I-SVM. The extended evaluation includes the AUC-ROC, CCR@FPR=1 %, CCR@FPR=10 %, and the ROC at a specific mixup-to-known ratio. The latter represents the point of optimal performance while exploiting mixups. The optimal point is indicated in the respective subtitle of each model.

In Fig. A.2a, OSNN demonstrates a failure case, exhibiting a degradation in performance across all metrics. However, we demonstrate that resolving the occupation problem enhances OSRs performance for this classifier.

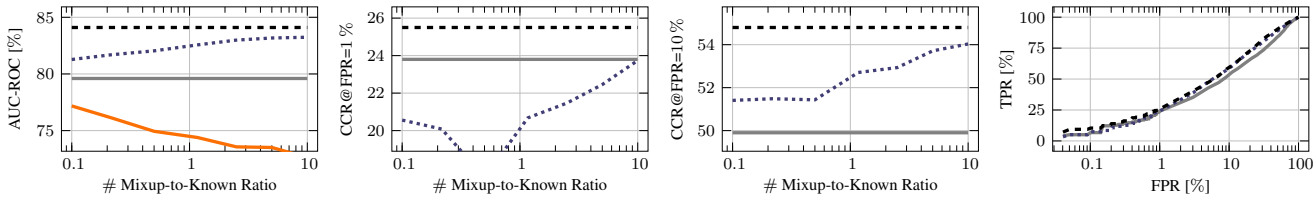
In Fig. A.2b, KvR demonstrates an improvement in the AUC-ROC, but this improvement comes partly at the expense of the CCR@FPR=1 %, which decreases to the mixup-to-known ratio of 0.3 and then starts to increases again. In comparison, CCR@FPR=10 % steadily increases and reaches 54 % at a mixup-to-known ratio of 10, while the baseline achieves 50 %. This gain is also evident in the ROC where it extends to very high FPRs. In conclusion, mixup proves to be a suitable approach for improving DNN (KvR) in applications with medium security requirements.

In Figs. A.2c and A.2d, the EVM and C-EVM outperform the baseline by exploiting mixups. Both variants show similar behavior, with KvR outperforming the other strategies in this experiment. The most significant improvement over the baseline is observed at the CCR@FPR=1 %, reaching over 25 % at the highest evaluated mixup-to-known ratio. This advantage also extends to the CCR@FPR=10 %, except for SPL, which experiences a sharp drop with many mixups, as previously observed in the AUC-ROC. The ROC shows that for SPL and MPL, the TPR deteriorates at FPRs greater than 10 %. This suggests that mixups have no positive effect on the closed-set case. However, there is a tremendous gain at FPRs between 0.1 to 10 %.

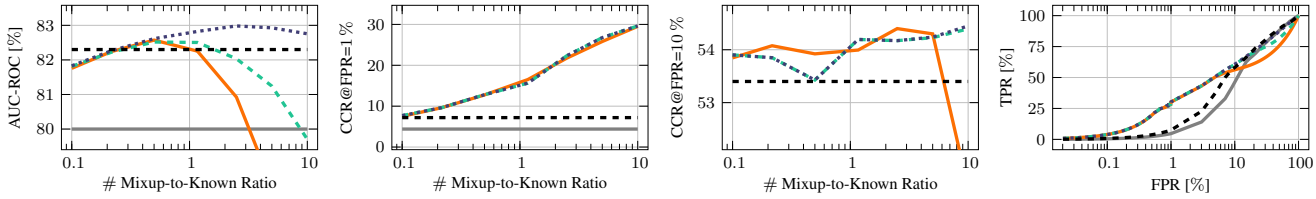
While the W-SVM in Fig. A.2e temporarily outperforms the baseline, the P_I-SVM in Fig. A.2f does not benefit from the mixup samples. It is noteworthy that training with genuine KUCs also did not results in any enhancement. A reasonable attempt at improvement would involve conducting a hyperparameter search for training with KUCs as well.



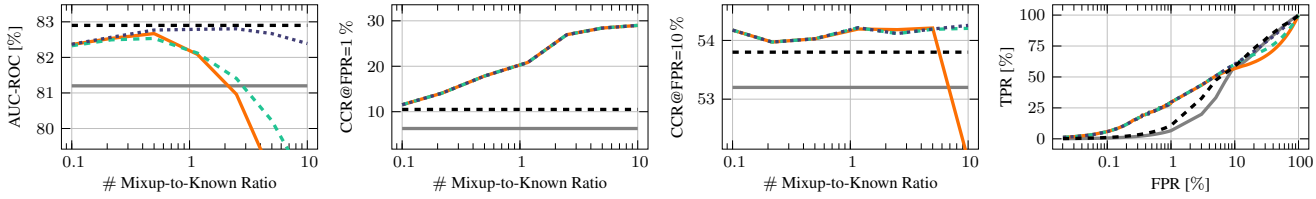
(a) OSNN, best = SPL, ROC is unavailable (baseline is not outperformed)



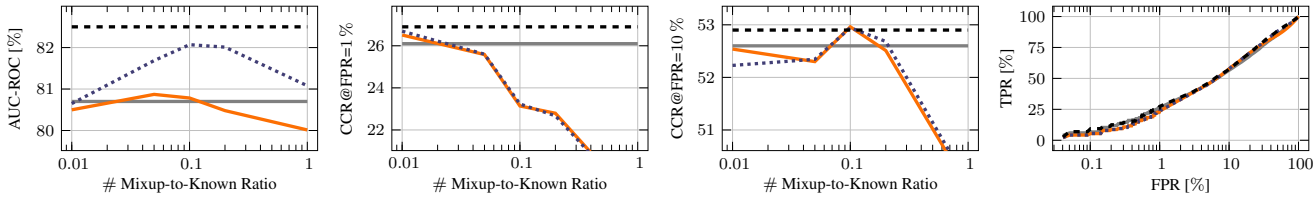
(b) DNN, best = KvR, ROC (righmost plot) is at mixup-to-known ratio = 10 with excluded DNN (SPL)



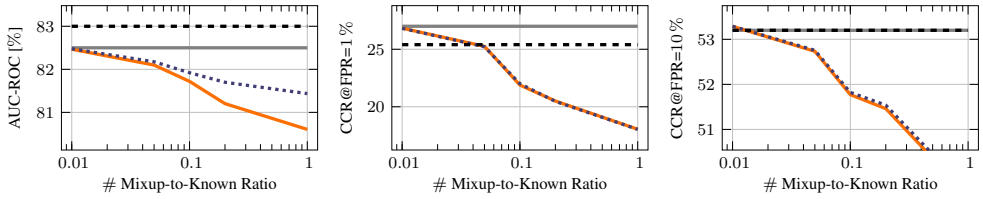
(c) EVM, best = SPL, ROC (rightmost plot) is at mixup-to-known ratio = 10



(d) C-EVM, best = SPL, ROC (rightmost plot) is at mixup-to-known ratio = 10



(e) W-SVM, best = KvR, ROC (rightmost plot) is at mixup-to-known ratio = 0.1



(f) P_I-SVM, best = KvR, ROC is unavailable (baseline is not outperformed)

Figure A.2. Unbiased results of all models (row-wise) trained with the different strategies and mixup samples on CIFAR-100. The metrics are (left to right): AUC-ROC, CCR@FPR=1 %, CCR@FPR=10 %, and the ROC. The first 3 metrics are shown w. r. t. the mixup-to-known ratio. The ROC is depicted at a specific mixup-to-known ratio. The baseline model without KUCs (—) and the best strategy of each model trained with genuine KUCs (- - -) from the first unbiased experiment, *cf*. Fig. 4 in the main work, serve as reference. This best strategy and the mixup-to-known ratio of the ROC are indicated in the respective subtitles.

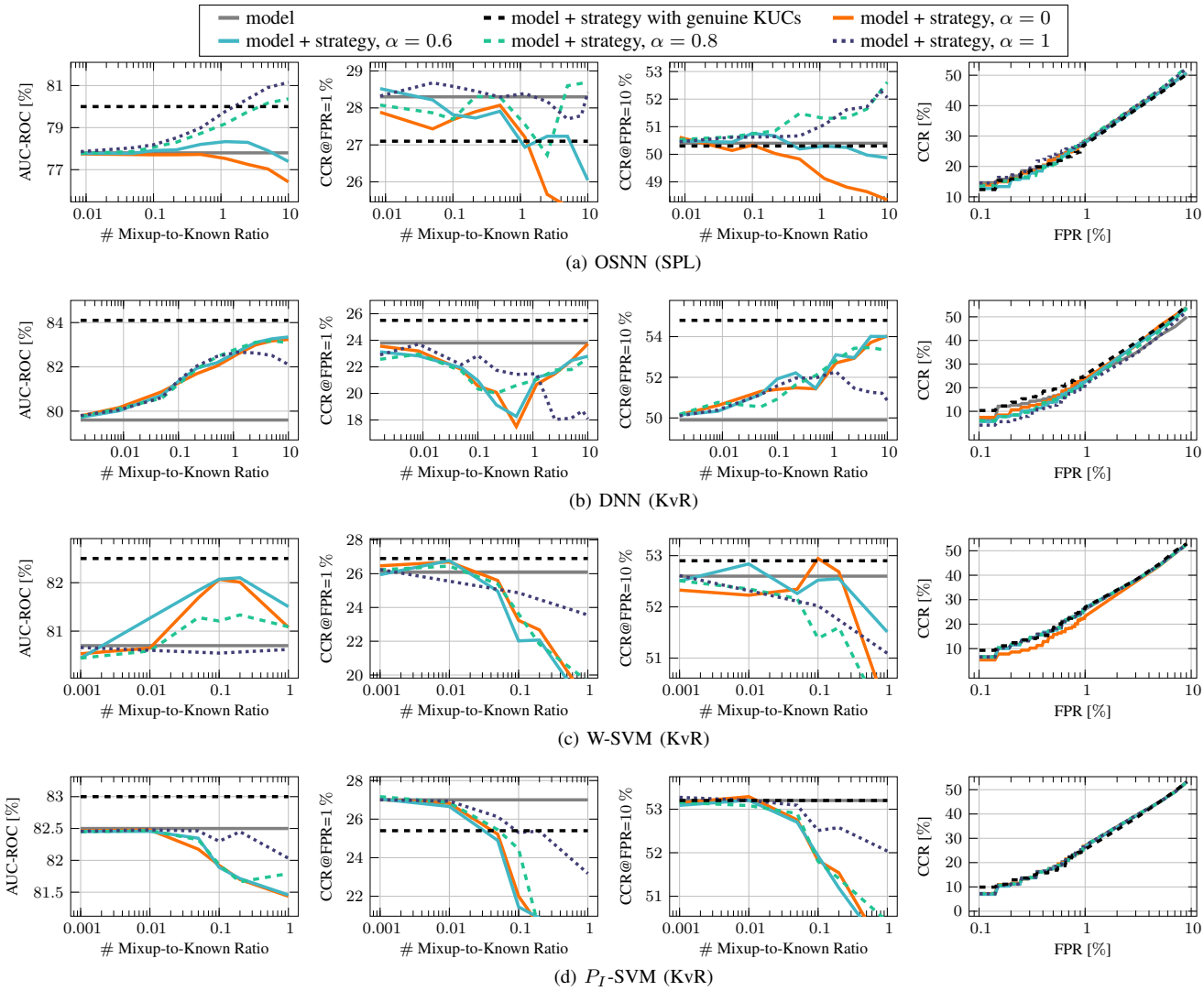


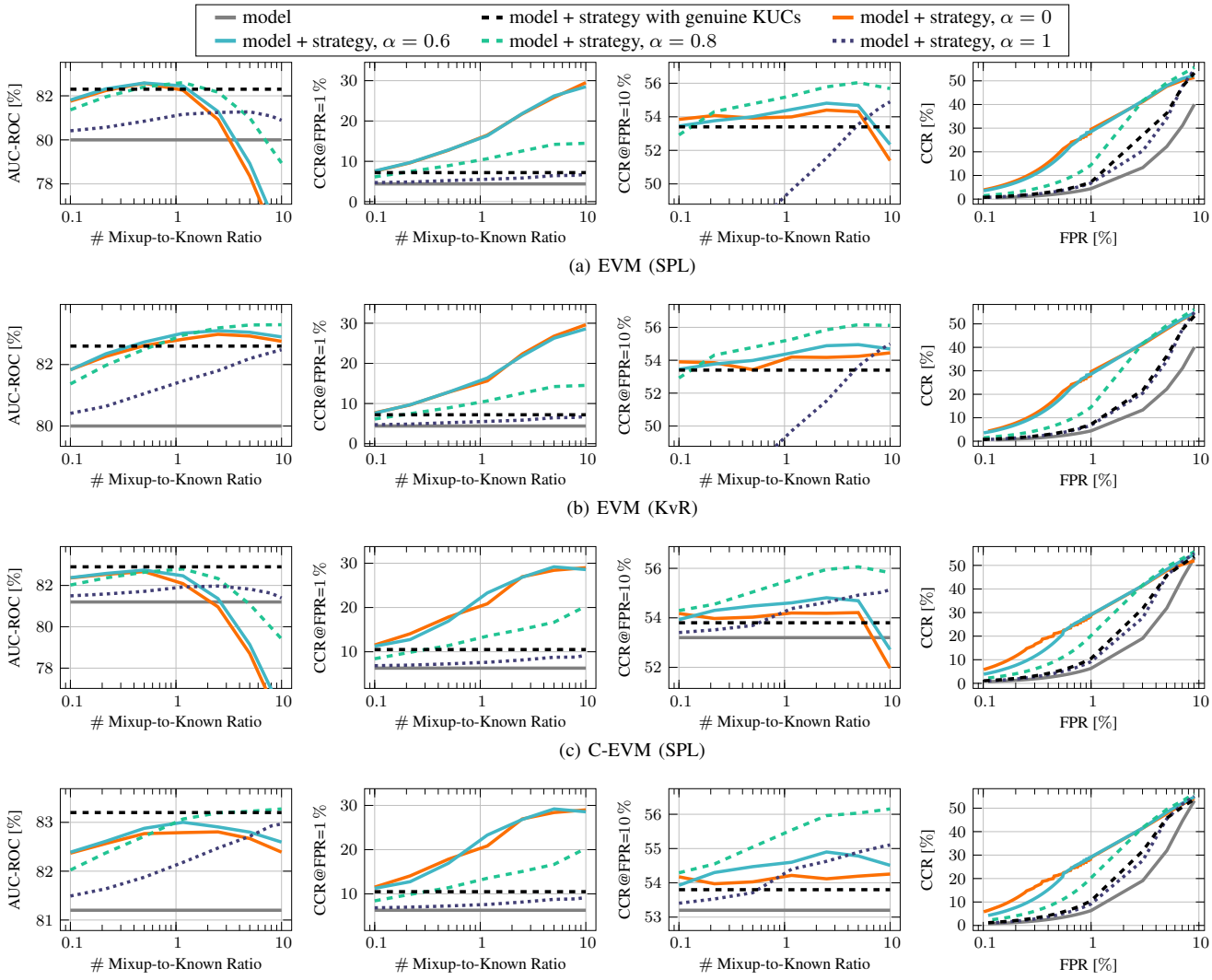
Figure A.3. Unbiased results of 4 models (row-wise) exploiting constrained mixups on CIFAR-100. The metrics are (left to right): AUC-ROC, CCR@FPR=1 %, CCR@FPR=10 %, and the OSCR at a certain mixup-to-known ratio. The OSCR corresponds to the maximum of each curve in the CCR@FPR=10 %. For example, in (a), for $\alpha = 0.8$ it is the mixup-to-known ratio of 10 and for $\alpha = 1$ the ratio of 8.

A.4. Solving the occupation problem – additional results

This section contains additional results of the assessment with constrained mixups to solve the occupation problem. While the main manuscript focuses on the AUC-ROC, here we provide the other open-set measures as well. Figure A.3 displays the results for the OSNN (SPL), DNN (KvR), W-SVM (KvR), and P_I-SVM. Figure A.4 contains the results for the EVM and C-EVM, both with SPL and KvR.

In general, the OSNN in Fig. A.3a benefits from more constrained mixups. While the CCR@FPR=1 % varies unstably, the CCR@FPR=10 % shows improvement. In contrast, the DNN (KvR) in Fig. A.3b does not show any improvements with constrained mixups. Stronger constraints can reduce the performance drop in the CCR@FPR=1 % by 3 %, but with $\alpha = 1$ it appears merely shifted. Both SVM variants in Figs. A.3c and A.3d marginally benefit from stronger constraints. Their overall downward trend is reduced with $\alpha = 1$ and could potentially be further improved with even stronger constraints. However, based on the current results, the exploitation of mixup, or KUCs in general, in combination with SVMs is limited for the detection of UUCs.

The EVM variants in Fig. A.4 exhibit minimal variation. Lower constraints enhance the CCR@FPR=1 % while a constraint with $\alpha = 0.8$ promotes the CCR@FPR=10 %. This difference in behavior can be leveraged when considering different safety requirements focused on different FPRs.



(a) EVM (SPL)

(b) EVM (KvR)

(c) C-EVM (SPL)

(d) C-EVM (KvR)

Figure A.4. Unbiased results of the EVM variants with SPL and KvR (row-wise) exploiting constrained mixups on CIFAR-100. The metrics are (left to right): AUC-ROC, CCR@FPR=1 %, CCR@FPR=10 %, and the OSCR at a specific mixup-to-known ratio. The OSCR corresponds always to the maximum of each curve at the CCR@FPR=1 %.