Knowledge-Distillation-Based Label Smoothing for Fine-Grained Open-Set Vehicle Recognition

Stefan Wolf[1,2] Dennis Loran[1,2] Jürgen Beyerer[2,1,3]

[1]Vision and Fusion Lab (IES)
Karlsruhe Institute of Technology
Karlsruhe, Germany

[2]Fraunhofer IOSB
Karlsruhe, Germany

[3]Fraunhofer Center
for Machine Learning
Munich, Germany

firstname.lastname@iosb.fraunhofer.de

Abstract

Fine-grained vehicle classification describes the task of estimating the make and the model of a vehicle based on an image. It provides a useful tool for security authorities to find suspects in surveillance cameras. However, most research about fine-grained vehicle classification is only focused on a closed-set scenario which considers all possible classes to be included in the training. This is not realistic for real-world surveillance applications where the images fed into the classifier can be of arbitrary vehicle models and the large number of commercially available vehicle models renders learning all models impossible. Thus, we investigate fine-grained vehicle classification in an open-set recognition scenario which includes unknown vehicle models in the test set and expects these samples to be rejected. Our experiments highlight the importance of label smoothing for open-set recognition performance. Nonetheless, it lacks recognizing the different semantic distances between vehicle models which result in largely different confusion probabilities. Thus, we propose a knowledge-distillation-based label smoothing approach which considers these different semantic similarities and thus, improves the closed-set classification as well as the open-set recognition performance.

1. Introduction

Searching getaway vehicles in surveillance cameras is a common task for security authorities after serious criminal offenses. While license plates provide a high specificity for searching vehicles and are easily identifiable, they usually get replaced before criminals commit an offense or the license plate number is unknown due to eyewitnesses usually not remembering them well. In these cases, fine-grained vehicle classification can be used to narrow down the possi-

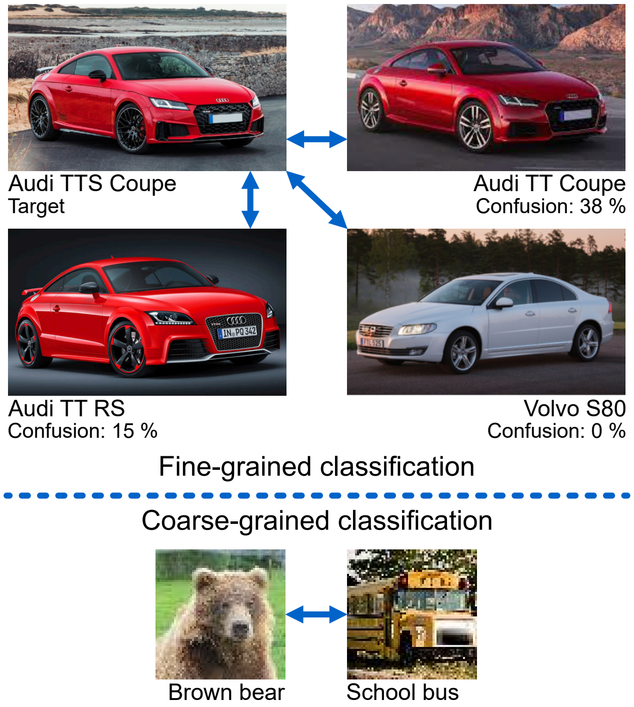


Figure 1. Top part: four different classes from the CompCars Web [48] dataset. In fine-grained vehicle classification, the semantic and visual difference between pairs of classes can vary significantly. However, this is not considered by the regular cross-entropy loss or by the label smooth loss. Our knowledge-distillation-based label smoothing considers the semantic difference between classes to regularize the training procedure and to reduce overconfidence of the network. The approximated confusion probabilities show that assigning the same confidence to all negative classes results in a harmed training process. In comparison, coarse-grained open-set recognition as symbolized on the lower part of the image with examples from the widely used Tiny-ImageNet [19, 34] is not affected to the same degree due to the variation of semantic distances between classes being lower compared to the overall distances between classes.

ble search space significantly since eyewitnesses remember the model of the vehicle easier than a license plate number. Thus, fine-grained vehicle classification is a widely investigated topic. However, usually a closed-set scenario is assumed in research while performing automatic vehicle search in real-world surveillance scenarios requires identifying unknown vehicle models due to the high number of available vehicle models. Training a model that supports all vehicle models which can occur in surveillance scenarios is practically impossible. Thus, in real-world use-cases, we have to handle an open-set scenario which is rarely investigated for fine-grained classification, particularly for fine-grained vehicle classification. Thus, we investigate the task of fine-grained open-set recognition for fine-grained vehicle classification.

Fine-grained object classification describes the task of image-based recognition of the class of an object on a fine-grained level with all possible classes sharing a common meta type. *E.g.*, recognizing the make and model of a car for fine-grained vehicle classification. One major challenge of this task is the large number of possible classes. Thus, for most fine-grained object recognition tasks it is impossible to learn all possible classes. Such a scenario is called open-set recognition which includes the usual task of distinguishing known classes but additionally requires the evaluated algorithm to recognize whether a sample is from a known class or from a class that was not part of the training set. The open-set recognition task is particularly difficult for fine-grained classification tasks due to only subtle differences distinguishing classes as can be seen in Fig. 1 which shows examples from the CompCars Web [48] dataset compared to examples from TinyImageNet [19]. This is further complicated by the fact that the semantic difference between classes can vary greatly. *E.g.*, an Audi TTS Coupe and an Audi TT Coupe are very difficult to distinguish due to only minor differences separating both while an Audi TT RS has a well distinguishable front end. Moreover, a Volvo S80 looks drastically different compared to the other vehicles and a confusion is highly unlikely. This highlights the different semantic distances between classes. However, as deep-learning models are traditionally trained with one-hot encoding, all negative classes are forced to have a zero confidence prediction. This is inappropriate for handling fine-grained vehicle classification in an open-set scenario since it leads to poorly-calibrated overconfident models due to not considering the semantic differences between classes. And, since calibration is highly important due to open-set recognition usually being based on the confidence of the model, this training procedure renders open-set recognition rather difficult. This problem is partially addressed by label smoothing [38] which adjusts the target labels by slightly reducing the target confidence of the positive class by a certain value α and increasing the target confidence of negative classes by a value of $1.0 - \frac{\alpha}{C-1}$ with C being the number of total classes in training. This leads to the confidence of the positive class not being fully maximized and the confidences of the negative classes not being fully minimized which better represents the true uncertainty of the model. Because of limitations of the data (*e.g.*, limited variety in terms of rims and vehicle colors [46]), a model will never be able to predict a class with 100% true certainty. This insight should be considered during training.

We extend the idea of label smoothing by not just using a fixed smoothing value α but instead incorporating class confusion probabilities based on knowledge-distillation in order to model the real visual and semantic distance between classes. Incorporating class confusion improves the model calibration since it enables forcing the model to be confident in cases where a confusion with other classes is unlikely and forcing the model to have a lower confidence when a confusion with another class in the training set is likely due to them being highly similar. This idea is applied to the smoothing of the label of the positive class as well as the negative classes which are all adaptively adjusted based on the likeliness of confusion. As a result, the better model calibration improves the open-set recognition performance while the reduction of overconfidence-induced overfitting improves the classification performance. This leads to target labels being closer to the true confusion probabilities. We illustrate this concept based on approximated confusion probabilities in Fig. 1. Since this advancements only adjusts the training loss, it does not increase the number of parameters of the model or its complexity during inference.

Our contributions can be summarized by: (1) we propose a knowledge-distillation-based label smoothing to improve the performance of classifiers for open-set performance without increasing the number of parameters or runtime complexity, (2) we provide extensive experiments on the impact of label smoothing on the open-set recognition performance and show the importance of choosing its α parameter properly and (3) we investigate the impact of the number of known classes on the open-set and closed-set task difficulty, thereby contradicting the common assumption that a higher openness [35] leads to a higher difficulty [41] and showing that increasing the number of classes is only slightly impacting the closed-set classification difficulty.

2. Related Work

Open-set recognition. Approaches for open-set recognition can be roughly divided into two categories. The **disrciminative-based approaches** can be further divided into softmax-based, distance-based, post-process-based and norm-based approaches. The softmax-based approaches use the confidence approximation by the softmax output of deep learning networks [14]. However, they tend to suffer from poor calibration limiting the open-set classification performance [12]. Nonetheless, the more recent Vision Transformer [7] architecture pre-trained on ImageNet21k has shown better calibration [10]. Distance-based approaches use the distance of features extracted from the input image to reference feature for predicting out-of-distribution samples. Different distance metrics have been used for this purpose [3, 20]. Ren et al. [33] propose a new relative Mahalanobis distance specifically tailored for distance-based open-set recognition. Miller et al. [28] propose a new loss that encourages a stronger separation of classes in feature space to improve open-set recognition performance. Post-process-based approaches aim to adjust the softmax scores without changing the model to improve the calibration. A commonly applied approach is ODIN [21] which uses temperature scaling and input perturbation. ODIN has been extended by GradNorm [17] by utilizing gradients explicitly and Generalized ODIN [15] by decomposing confidence scores and modifying input preprocessing. Liu et al. [22] propose an energy score as an alternative post-process optimization of confidence scores. Norm-based approaches utilize the norm of feature vectors or logits from the network prior to the softmax activation. Vaze et al. [41] found that the magnitude of the logits provide a better metric for open-set recognition than softmax scores. Wei et al. [43] normalize the vector norm of the logits during training by an additional loss to reduce the overconfidence of the network. **Generative-based approaches** generate new samples synthetically that support learning distinguishing kn won from unknown samples. Ge et al. [11] propose G-OpenMax which uses a GAN to generate samples that are targeted to be of novel classes by exploiting interpolation in the latent space of the GAN. C2AE [29] uses an auto-encoder to perform open-set recognition based on the reconstruction error of the query sample assuming that the reconstruction capability of unknown classes is worse than for known classes.

Fine-grained open-set recognition. While a wide range of research was done for open-set recognition, they mostly focus on coarse-grained tasks with a high semantic difference between classes. Fine-grained open-set recognition, *e.g*. performing open-set recognition for fine-grained vehicle classification, has been rarely investigated. Vaze et al. [42] propose the semantic shift benchmark for evaluating open-set recognition methods which incorporates three fine-grained classification datasets for birds, aircrafts and cars. However, each of these datasets do not contain more than 200 classes, rendering it a relatively easy task. Other studies in the field suffer from similar issues due to using the same datasets for evaluation [37, 52]. To the best of our knowledge, more complicated fine-grained open-set recognition tasks have yet only been evaluated in the field of fungi classification [9, 16, 30–32, 44, 45, 49].

3. Methodology

In this section, we first introduce the open-set scenario formally and then, describe our classification architecture with its open-set recognition strategy and its training procedure. Finally, we describe our novel knowledge-distillation-based label smoothing approach.

Open-set scenario. In an open-set scenario, the total classes can be divided into known classes $\mathcal{C}$ with $C = |\mathcal{C}|$ and unknown classes $\mathcal{U}$. Based on this distinction, the training dataset $\mathcal{D}_{\text{train}}$ and testing dataset $\mathcal{D}_{\text{test}}$ can be split into two subsets each: the subset of samples from known classes $\mathcal{D}_{s\text{-known}} = \{(\mathbf{X}, y) \in \mathcal{D}_s | y \in \mathcal{C}\}$ and the subset of samples from unknown classes $\mathcal{D}_{s\text{-unknown}} = \{(\mathbf{X}, y) \in \mathcal{D}_s | y \in \mathcal{U}\}$ for each $s \in \{\text{train}, \text{test}\}$. During training, only the subset of samples of known classes $\mathcal{D}_{\text{train-known}}$ is used and the samples from unknown classes $\mathcal{D}_{\text{train-unknown}}$ are ignored. Afterwards, during evaluation, the classification algorithm has to perform two tasks. The first task is to identify whether a sample belongs to the known classes $\mathcal{C}$ or the unknown classes $\mathcal{U}$ on the complete test set $\mathcal{D}_{\text{test}}$. The second task is the usual closed-set classification, *i.e*. distinguishing the known classes $\mathcal{C}$ on there corresponding subset of the training set $\mathcal{D}_{\text{test-known}}$. A good open-set recognition algorithm should perform well on both tasks.

Architecture. We build our classification architecture based on a modern Swin Transformer V2 [23] feature extraction backbone combined with a linear classification layer and a softmax activation function. The linear classifier outputs a logit vector with each element representing an unnormalized confidence of the class. Afterwards, the softmax activation function is applied on the logit vector to normalize the cofidence scores. So, the final output of the network is a vector of confidence scores between 0 and 1.

Open-set recognition. To perform an open-set recognition, we have to perform a class-wise out-of-distribution detection. To achieve this, we use the softmax confidences from the classification network. We identify a sample as an out-of-distribution sample when the maximal softmax score is below a certain threshold. We do not specify this threshold for the experiments since the chosen metrics adjust the threshold to compromise recall and precision of the out-of-distribution detection. For deployment, a certain threshold has to be chosen as a specific working point that achieves a compromise well-suited for the task at hand. This can be a higher focus on recall or a higher focus on precision.

Training procedure. The deep learning network is trained by assigning a target vector and minimizing the difference of the softmax layer output to the target vector by using a cross-entropy loss function. The optimization is done by performing back propagation with the AdamW [26] algorithm. Usually, the target vector $\mathbf{p_0}(\mathbf{X})$ is built by using a one-hot encoding as follows:

$$p_{0,i}(\mathbf{X}) = \begin{cases} 1 \text{ if } i = y \\ 0 \text{ else} \end{cases}, \tag{1}$$

with y being the target class of the input sample $\mathbf{X}$.

Label smoothing. Our label assignment strategy is based on label smoothing as proposed by Szegedy et al. [38]. Label smoothing adjusts the one-hot encoded label by smoothing between the positive class and the negative classes. *I.e.*, the generated target vector has a value of $1 - \alpha$ for the element of the ground truth class while the elements of all other classes have a value of $\frac{\alpha}{C-1}$ with α being a hyper parameter and C being the total number of known classes. This means that for a given one-hot encoded target vector $\mathbf{p_0}(\mathbf{X})$, the label smoothed target vector can be calculated by

$$\mathbf{p_0}(\mathbf{X})^{\text{LS}} = \mathbf{p_0}(\mathbf{X}) \cdot (1 - \alpha) + (\mathbf{1} - \mathbf{p_0}(\mathbf{X})) \cdot \frac{\alpha}{C-1}. \tag{2}$$

Label smoothing regularizes the training by not forcing overconfident predictions due to the smoothed target labels not having the highest possible confidence. This in turn results in a more calibrated model that achieves a better open-set recognition performance as shown in Sec. 4.

Knowledge-distillation-based label smoothing. Label smoothing is assigning the same ground-truth confidence to each non-negative softmaxed logit. This should prevent overfitting since it prevents the training process from incentivizing overconfidence and thus, works as a method of regularizing the network. However, for open-set recognition, a well-calibrated model is highly important to have the model predict confidences which correspond to the actual accuracy of the model. In this regard, label smoothing lacks taking the differences between classes into account. Particularly for fine-grained classification, the semantic difference between classes can be largely different. *E.g.*, for fine-grained vehicle classification, a confusion with a vehicle model of the same make is more likely than with a model of a different make which should be considered in the label assignment process. Thus, we propose a knowledge-distillation-based label smoothing technique that assigns class-confusion-aware labels.

First on, we need to find the probabilities of the confusions. These should be in the form of a confusion matrix $\mathbf{\Sigma} \in \mathbb{R}^{C \times C}$. Each entry $\Sigma_{i,j}$ describes the probability that class i is confused with class j. These entries could be approximated by the hierarchy of the dataset, *e.g.*, similarity in terms of make or other attributes. However, we propose a more fine-grained approach based on knowledge distillation. We first train a model without label smoothing that we use as teacher network to predict the probabilities of class confusions. We evaluate the model on the train set and extract the logits before the softmax activation function. The logit of the ground-truth label is then set to zero as we do not want to adjust for confusion with the class itself. Afterwards, the adjusted logits are normalized by a softmax function and then averaged per class logit for each ground-truth class. This results in the entries of our confusion matrix $\mathbf{\Sigma}$.

Thereafter, we train our final model as a student network using the confusion matrix as distilled knowledge from the teacher network. Let $\mathbf{p_0}(\mathbf{X})$ the one-hot encoded ground-truth label. The student network get the target label

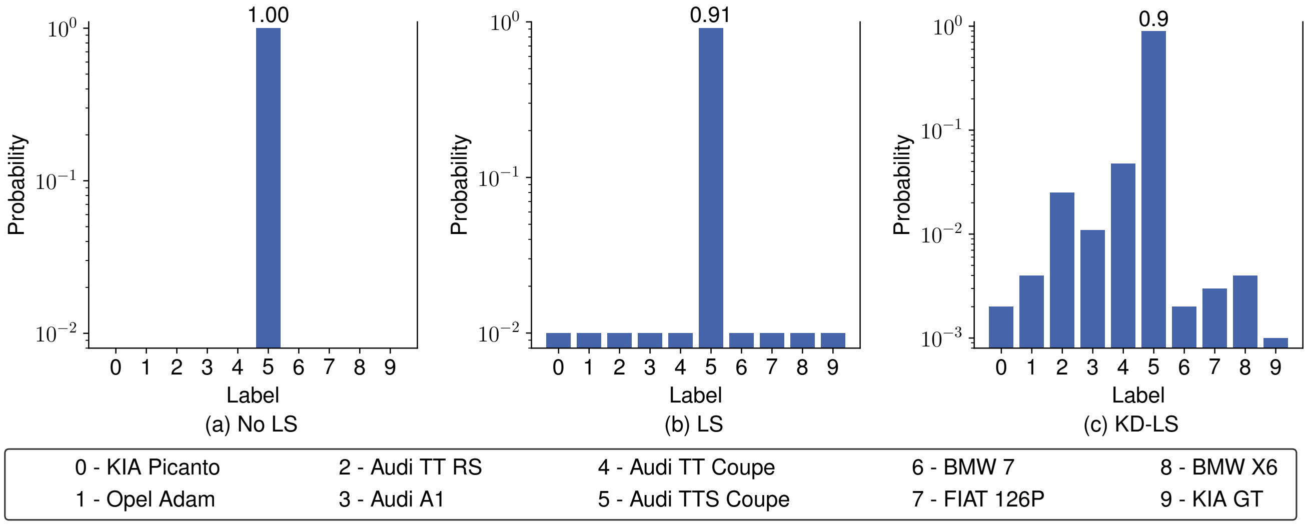
$$\mathbf{p_0}(\mathbf{X})^{\text{KD-LS}} = \mathbf{p_0}(\mathbf{X}) \cdot (1 - \alpha) + \mathbf{\Sigma}_k \cdot \alpha \tag{3}$$

assigned with $\mathbf{\Sigma}_k$ being the k-th row of $\mathbf{\Sigma}$ which describes the probability of confusion of class k with all other classes and α being the smoothing value as also applied for label smoothing. The impact of our knowledge-distillation-based label smoothing on the target vector is illustrated in Fig. 2.

4. Evaluation

Dataset. For training and evaluation, we mainly use the CompCars Web [48] dataset with a total of 136,725 images. While the dataset contains annotations regarding the make, model and year of the vehicle for each image, we perform the classification on the level of the model to prevent that some classes have too few images. Thus, in our setup, the dataset has 1,716 classes. CompCars Web contains a high diversity of images in terms of vehicle poses and photo backgrounds. However, as Buzelle and Segation [4] and Wolf et al. [46] highlighted, the default random train-test-split is heavily biased leading to a split which can be easily solved. Thus, we use a harder train-test-split which is based on maximizing the feature distance of a ResNet-18 feature extractor between the train and the test set [4]. The split has a ratio of 70% train and 30% test samples. Since we evaluate an open-set scenario, we split the total of 1,716 classes into 858 known and 858 unknown classes. For training, only the training images of the known classes are used. For evaluation, metrics involving on the classification performance are only evaluated on the images of the test set of the known classes while metrics involving open-set recognition performance also use the images of the unknown classes from the test set. We choose 20% of the training set randomly as validation set to choose the best model from each training run. However, all reported metrics are from evaluations on the test set.

We verify the effectiveness of our approach on the CompCars Surveillance (SV) dataset [48]. CompCars SV cotains 44,481 images with a total of 281 classes. The preprocessing is similar to CompCars Web. For the SV experiments, we generate a harder train-test-split with the method



(a) No LS

(b) LS

(c) KD-LS

Figure 2. Comparing different target label assignment strategies: traditional one-hot encoding (a), label smoothing (b) and our knowledge-distillation-based label smoothing (c). Label smoothing prevents overfitting by not incentivizing overconfidence. However, assigning the same label to all negative classes is inappropriate for modeling confidences of fine-grained classification where the semantic distance between classes can vary significantly. Our knowledge-distillation-based label smoothing takes the fine-grained classification setting into account to assign labels with a proper modeling of the probabilities of confusion.

of Buzelle and Segation [4] that gives a better approximation of real-world performance. The split has 70% train and 30% test ratio with 20% of the train samples used for validation. 141 of the 281 vehicle models are used as known classes with the remaining 140 being the unknown classes for which no samples are available during training.

Metrics. To evaluate the classification performance, we use the macro-averaged F1-score which is evaluated on the images of the known classes of the test set. To evaluate the open-set recognition performance, we use the *area under the receiver operating characteristic* (AUROC) metric which we evaluate on images of the test set as binary classification problem to distinguish the images of known from images of the unknown classes. We also report the *open-set classification rate* (OSCR) [6] which indicates both the classification and the open-set recognition performance.

Evaluation setup. We train the models with AdamW [26], a batch size of 32, an initial learning rate of 10^{-3}, a weight decay of 0.05 and a cosine annealing learning rate schedule that reduces to learning rate to 1% of the initial learning rate. We start the training with a warm-up phase of 2 epochs with a linear learning rate increase from $10^{-3}\times$ to $1.0\times$ of the initial learning rate. The training is run for 200 epochs in total with the best checkpoint being selected for testing based on the F1 score on the validation set. All models are pre-trained on the ImageNet1k [34] dataset.

During training, a random crop with the size of 8% to 100% of the original image is taken and resized to the network input size. Afterwards, a random horizontal flip and RandAugment [5] are applied with RandAugment using the default set of policies but only one policy per sample, a total level of 14, a magnitude level of 10 and no magnitude deviation. Preliminary experiments have shown these settings to be advantageous compared to the default settings. Finally, the augmented image is normalized by the mean and standard deviation of CompCars Web. During evaluation, the image is first resized to 114% of the network input size on the shorter side and afterwards, a center crop with the network input size is taken. Finally, the image is normalized similar to the training. The network input size is 256×256 for Swin Transformer V2 and 224×224 for all other backbones due to architectural particularities of Swin Transformer V2. We refrain from employing CutMix [50] or *mixup* [51] data augmentations even though they are widely applied in recent training recipes [23–25, 40] since preliminary experiments have shown that it decreases the open-set recognition performance which could not outweigh gains in closed-set accuracy.

State-of-the-art open-set recognition. We compare our knowledge-distillation-based label smoothing to recently published improvements for open-set recognition like maximum logit score (MLS) [41] and out-of-distribution-detection like energy score [22] and LogitNorm [43] on the CompCars Web and the CompCars SV datasets. For a fair comparison, all methods are evaluated with our tuned baseline using a Swin Transformer V2 Small model, optimized data augmentation and label smoothing with an adjusted α

Method	Dataset					
	CompCars Web			CompCars Surveillance		
	F1	AUROC	OSCR	F1	AUROC	OSCR
Baseline	92.3	93.9	90.8	91.9	94.4	90.6
Maximum Logit Score	92.3	**94.1**	90.7	91.9	93.5	89.7
Energy-Score	92.3	93.8	90.3	91.9	93.2	89.4
LogitNorm ($\tau = 0.05$)	92.1	87.3	84.6	**92.9**	94.6	90.7
Knowledge-Distillation-Based Label Smoothing (ours)	**93.3**	94.0	**91.2**	92.1	**95.0**	**91.1**

Table 1. Comparing our knowledge-distillation-based label smoothing approach to other recently published approaches for open-set recognition or model calibration in general. For a fair comparison, all approaches are applied to the same baseline including. For CompCars Web the other evaluated approaches provide no advantage on top of the baseline apart from a slight increase in terms of AUROC with maximum logit score. Our approach increases the performance, particularly the F1 score, significantly. On CompCars Surveillance, our approach also provides an advantage over the baseline for all evaluated metrics.

value. Since maximum logit score and the energy score are just adapting the post-training open-set recognition, the F1 score does not change compared to the baseline. Regarding the open-set recognition performance as measured by the AUROC score, MLS leads to a slight increase while the energy score leads to a slight decrease for the CompCars Web. However, both show a drop in terms of the mixed performance as measured by the OSCR. On the CompCars SV dataset, both methods show a significant drop in terms of both AUROC and OSCR. The LogitNorm loss leads to a slight reduction of closed-set accuracy and a large reduction in terms of open-set recognition performance and in turn, leads to a strongly reduced OSCR score on the CompCars Web. However, on the CompCars SV dataset, the logit normalization can significantly improve the F1 score and shows a slight increase in terms of open-set recognition performance in terms of AUROC and thus, also improves the OSCR slightly. Our KD-LS approach shows an improvement for all metrics with the largest one being on the F1 score for the CompCars Web. Compared to the other approaches, only the MLS can slightly outperform KD-LS on the CompCars Web regarding the AUROC score and the LogitNorm can outperform KD-LS on the CompCars SV regarding the F1 score. Nonetheless, KD-LS shows a high consistency with not falling behind heavily in any of the metrics or datasets evaluated. Finding the exact reason for partially strong drops of the other methods, *e.g.* LogitNorm dropping by about 6 percentage points regarding OSCR on CompCars Web compared to the baseline, is up to future research. It might be related due to LogitNorm being more sensitive to hyper parameters. The drops of MLS and energy score in regards to the AUROC and OSCR scores on the CompCars SV seem to be specific to this dataset. However, particularly for open-set recognition, consistency is highly important due to the openness of the task rendering sophisticated evaluation on real-world scenarios difficult.

Method	α	F1	AUROC	OSCR
Baseline	-	92.6	91.2	88.8
Label Smoothing	10^{-1}	91.9	82.3	80.2
	10^{-2}	92.3	93.9	90.8
	10^{-3}	92.0	93.1	90.2
Knowledge-Distillation-Based Label Smoothing	10^{-1}	92.2	88.9	86.2
	10^{-2}	**93.3**	**94.0**	**91.2**
	10^{-3}	92.2	91.4	88.8

Table 2. Evaluating our knowledge-distillation-based label smoothing approach for open-set recognition on the fine-grained vehicle classification dataset CompCars Web. While label smoothing shows a strong increase in the AUROC score as indicator for the open-set recognition performance, it slightly degrades the F1 classification score. Our knowledge-distillation-based label smoothing approach is slightly increasing the AUROC score further but its main advantage is a significant increase for the F1 classification score which in turn leads to an increased combined OSCR score.

Knowledge-distillation-based label smoothing. We evaluate our knowledge-distillation-based label smoothing approach in more detail and show the results in Table 2. First on, we run evaluations on the CompCars Web dataset regarding static label smoothing which is indifferent to class confusion. A crucial hyper parameter for label smoothing is α which describes the confidence set as target for negative classes during training. A value of $\alpha = 0$ corresponds to not employing label smoothing at all. We compare no label smoothing to values of $\alpha \in \{10^{-1}, 10^{-2}, 10^{-3}\}$. While all evaluated values reduce the F1-Score indicating closed set classification performance compared to no label smoothing,

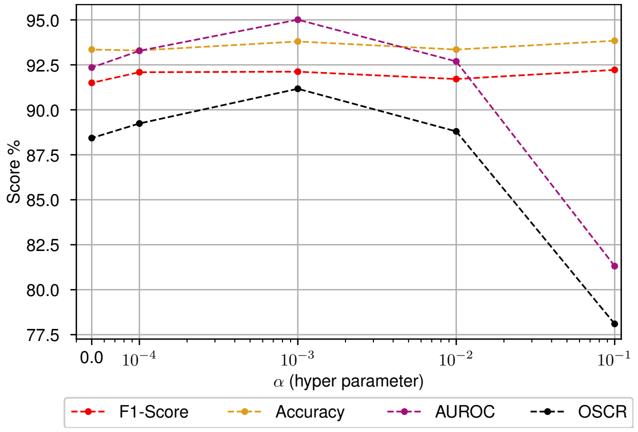


Figure 3. Evaluation of the different metrics on the CompCars Surveillance dataset for a model trained with knowledge-distillation-based label smoothing with different values for α. Label smoothing has a significant advantage for the classification and open-set recognition performance as long as the value of α is chosen carefully. Choosing a value too large can degrade the performance of the model beyond training without label smoothing.

the AUROC score measuring open-set recognition performance achieves its optimum at $\alpha = 10^{-2}$ with 2.7 points over the baseline without label smoothing. However, using 10^{-1} degrades the AUROC by 8.9 points compared to not using label smoothing. This highlights the impact of the value of α and the requirement to choose it carefully.

We analyze the impact of α for knowledge-distillation-based label smoothing in more detail for the CompCars SV dataset in Figure 3. It can be seen that the impact rises with an in creasing α value as expected. A value of 10^{-4} results in only a slight increase in terms of F1, AUROC and OSCR. With a value of 10^{-3} the results clearly peak for the AUROC and the OSCR. Increasing α further brings these two metrics close to the initial value without any label smoothing at 10^{-2} and showing a heavy degradation with an even higher value of 10^{-1}. In contrast, the closed-set metrics F1 and accuracy do not show a clear tendency. F1 shows an increase when increasing α from 0 to 10^{-4} but only shows slight variations for higher values. This highlights the importance of properly choosing α for open-set recognition performance. Particularly, the optimal value is dependent on the dataset and thus, it should be chosen independently for each dataset.

Model architectures. Model architectures is a heavily researched field. A wide range of architectures have been invented which are either general-purpose architectures designed for overall better generalization or are designed for specific purposes like fine-grained classification. We evaluate different size variants of ResNet [13], ResNeXt [47], ConvNeXt [25], Swin Transformer [24] and Swin Transformer V2 [23] which are general-purpose architectures and

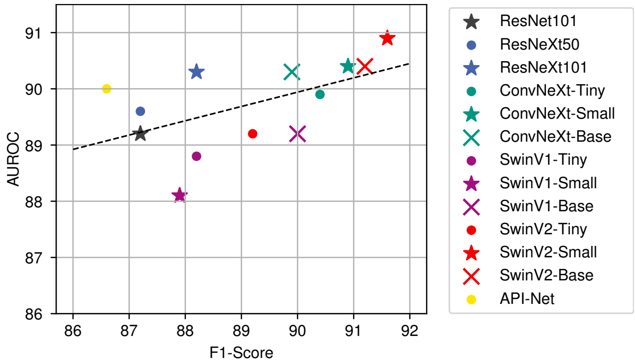
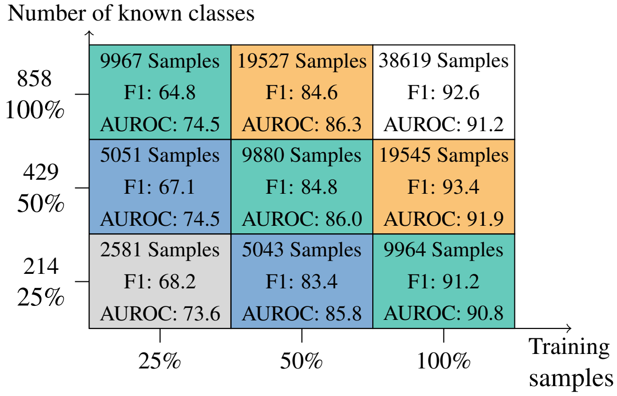


Figure 4. The figure shows different classification architecture with their open-set recognition performance in terms of AUROC compared to the classification performance in terms of F1 score.

we evaluate API-Net [53] as an architecture optimized for fine-grained classification. The models are evaluated with only basic data augmentation (horizontal flip) and without label smoothing. The results are shown in Figure 4. As expected, there is a clear tendency that more modern backbones are achieving a better F1 score. While larger model are mostly performing better than their smaller counterparts, at least for ConvNeXt and Swin V2, the Base variants are showing a worse performance than the Small variants. This is a sign of overfitting and more regularization like data augmentation or label smoothing might be needed to make the Base models perform better. Interestingly, there is a strong correlation between the F1 score and the AUROC score. These findings are in line with the findings by Vaze et al. [41] who state that the open-set performance is closely related to the closed-set performance. This shows that either better model calibration is necessary for better closed-set performance or vice versa. However, some models are off the average ratio between AUROC and F1 score. The Swin V1 models show an AUROC score which is significantly below what is to be expected based on its F1 score. And the API-Net, optimized for fine-grained classification, is significantly above the average ratio showing a high open-set performance compared to its closed-set performance. Nonetheless, due to it having the lowest F1-score in the comparison, it it not achieving the best AUROC score.

Impact of number of classes and samples. A wide range of datasets for fine-grained vehicle classification [2, 18, 27, 36, 39, 48] has been proposed with different number of classes and different sample sizes. However, there is a lack of systematic investigation regarding the impact of the number of classes and the sample sizes on the difficulty of the task and the achieved classification performance, particularly for open-set scenarios. Thus, we evaluate our approach for subsets of CompCars Web that are either reduced by the number of samples per class, the number of total known classes or both. We hypothesize that (1) a higher number



Number of known classes	25%	50%	100%
858 100%	9967 Samples F1: 64.8 AUROC: 74.5	19527 Samples F1: 84.6 AUROC: 86.3	38619 Samples F1: 92.6 AUROC: 91.2
429 50%	5051 Samples F1: 67.1 AUROC: 74.5	9880 Samples F1: 84.8 AUROC: 86.0	19545 Samples F1: 93.4 AUROC: 91.9
214 25%	2581 Samples F1: 68.2 AUROC: 73.6	5043 Samples F1: 83.4 AUROC: 85.8	9964 Samples F1: 91.2 AUROC: 90.8

Training samples

Figure 5. The figure shows results of evaluations executed on subsets of the CompCars Web. We reduce the number of classes and the number of samples per class systematically to investigate the impact of these two factors. In contrast to our expectation, a higher number of classes is not rendering the closed-set classification task significantly more difficult. By far more important is the number of samples per class. Also in contrast to our expectation, a higher number of classes is not increasing the open-set recognition performance. Again, most important is the number of classes per sample.

of classes leads to a worse closed-set classification performance for the same number of samples per class and (2) the open-set performance will improve for an increasing number of known classes for the same number of samples per class. We assume hypothesis (1) since more classes will reduce the probability of choosing the correct classes when only choosing randomly. So, a random guessing classifier will perform worse and a real algorithm has to achieve a higher relative performance compared to the random guessing classifier in order to achieve the same classification accuracy than with a lower number of classes. And without having more samples available per class, the ability to distinguish classes should not improve. This means that the task should get more difficult with more classes. We assume hypothesis (2) since more classes should provide a feature space that is generalizing better to distinguish any vehicle classes and so, also new classes. Thus, the new feature space should provide a better representation for distinguishing unknown classes from any of the known classes. Additionally, this hypothesis is supported by the openness [35] metric which assigns a task an openness score based on the ratio of the number of unknown classes to the number of known classes. The openness score is a common metric for estimating the difficulty of open-set recognition tasks.

For the evaluations, we use our baseline with a Swin Transformer V2 backbone and optimized data augmentation but without label smoothing. The results of the experiment are shown in Fig. 5. Regarding hypothesis (1), as expected, we see a drop in terms of F1 score when we increase the number of classes from 25% to 50% and to 100% of the total number of known classes when using 25% of the total number of samples per class. However, with 50% or 100% of the total samples, the differences in terms of F1 score between the different number of classes evaluated become insignificant considering that the classes and thus, the samples in the test set change between the evaluations. So, we can conclude that as long as the number of samples is not becoming too low, we can reject the hypothesis and the number of classes is not significant for the closed-set classification performance for the range of evaluated number of classes. This insight opens space for future research for the reason behind this unexpected behavior and the limits of the new hypothesis. The reason our hypothesis does not hold might be that the generalization of the total number of samples which increase with more classes as long as the number of samples per class is kept the same outweighs the increased difficulty. This is supported by the fact that when looking at the results with the total number of samples being the same while increasing the number of classes, *i.e.* the results with a common color in the figure, the F1 shows a drastic drop in performance.

Regarding hypothesis (2), the results do not reflect the expected behavior with the results showing an AUROC score being largely independent of the number of known classes in training considering that the train and test sets being not exactly the same due to the changing classes. This holds true as long as the number of samples per classes are kept the same. As soon as the number of samples per class increases, the AUROC score also increases significantly. A reason for the unexpected behavior might be the strong correlation of open-set recognition and closed-set classification performance with the last one also not changing significantly with the number of classes. The observed behavior contradicts the use of the openness metric [35] as a measure for open-set recognition difficulty.

5. Conclusion

We investigate the open-set recognition scenario in the context of fine-grained vehicle classification. This is particularly important for real-world surveillance applications due to them not being controllable in terms of vehicle models shown to the classification model. We show that recent state-of-the-art deep learning models already achieve a good performance for both the closed-set classification as well as the open-set recognition. Nonetheless, both can still be improved with our proposed knowledge-distillation-based label smoothing. We execute additional experiments which show that the drop in classification accuracy is insignificant with increasing number of classes as long as the number of samples can be kept the same. Additionally, we show that the number of samples per class is crucial for the open-set recognition performance while the number of classes has a negligible impact.

References

[1] Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos, editors. *Proceedings of the Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (CLEF)*, number 3497 in CEUR Workshop Proceedings, Aachen, 2023. 9, 10

[2] Abdollah Amirkhani and Amir Hossein Barshooi. Deepcar 5.0: Vehicle make and model recognition under challenging conditions. *IEEE Transactions on Intelligent Transportation Systems*, 24(1):541–553, 2023. 7

[3] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[4] Marco Buzzelli and Luca Segantin. Revisiting the compcars dataset for hierarchical car classification: New annotations, experiments, and results. *Sensors*, 21(2), 2021. 4, 5

[5] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 5

[6] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 5

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[8] Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast, editors. *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (CLEF)*, number 3180 in CEUR Workshop Proceedings, Aachen, 2022. 9, 10

[9] Gao Fan, Chen Zining, Wang Weiqiu, Song Yinan, Su Fei, Zhao Zhicheng, and Chen Hong. Does closed-set training generalize to open-set recognition? In Faggioli et al. [8], pages 2063–2077. 3

[10] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7068–7081. Curran Associates, Inc., 2021. 3

[11] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017. 3

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. 3

[13] 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)*, June 2016. 7

[14] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 2

[15] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[16] Feiran Hu, Peng Wang, Yangyang Li, Chenlong Duan, Zijian Zhu, Yong Li, and Xiu-Shen Wei. A deep learning based solution to fungiclef2023. In Aliannejadi et al. [1], pages 2051–2059. 3

[17] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 677–689. Curran Associates, Inc., 2021. 3

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 7

[19] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 1, 2

[20] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 3

[21] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 3

[22] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020. 3, 5

[23] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, June 2022. 3, 5, 7

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Con-*

ference on Computer Vision (ICCV), pages 10012–10022, October 2021. 5, 7

[25] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, June 2022. 5, 7

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 4, 5

[27] Lei Lu, Ping Wang, and Hua Huang. A large-scale frontal vehicle image dataset for fine-grained vehicle categorization. *IEEE Transactions on Intelligent Transportation Systems*, 23(3):1818–1828, 2022. 7

[28] 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, January 2021. 3

[29] 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)*, June 2019. 3

[30] Lukáš Picek, Milan Šulc, Rail Chamidullin, and Jiří Matas. Overview of fungiclef 2023: fungi recognition beyond 1/0 cost. In Aliannejadi et al. [1], pages 1943–1953. 3

[31] Lukáš Picek, Milan Šulc, Jiri Matas, and Jacob Heilmann-Clausen. Overview of fungiclef 2022: Fungi recognition as an open set classification problem. In Faggioli et al. [8], pages 1970–1981. 3

[32] Huan Ren, Han Jiang, Wang Luo, Meng Meng, and Tianzhu Zhang. Entropy-guided open-set fine-grained fungi recognition. In Aliannejadi et al. [1], pages 2122–2136. 3

[33] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021. 3

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1, 5

[35] 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*, 35(7):1757–1772, 2013. 2, 8

[36] Jakub Sochor, Jakub Špaňhel, and Adam Herout. Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance. *IEEE Transactions on Intelligent Transportation Systems*, 20(1):97–108, 2019. 7

[37] Jiayin Sun, Hong Wang, and Qiulei Dong. Spatial-temporal attention network for open-set fine-grained image recognition. *arXiv preprint arXiv:2211.13940*, 2022. 3

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 4

[39] Faezeh Tafazzoli, Hichem Frigui, and Keishin Nishiyama. A large and diverse dataset for improved vehicle make and model recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 7

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021. 5

[41] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022. 2, 3, 5, 7

[42] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. The semantic shift benchmark. In *ICML 2022 Shift Happens Workshop*, 2022. 3

[43] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23631–23644. PMLR, 17–23 Jul 2022. 3, 5

[44] Stefan Wolf and Jürgen Beyerer. Transformer-based fine-grained fungi classification in an open-set scenario. In Faggioli et al. [8], pages 2219–2226. 3

[45] Stefan Wolf and Jürgen Beyerer. Optimizing fine-grained fungi classification for diverse application-oriented open-set metrics. In Aliannejadi et al. [1], pages 2159–2167. 3

[46] Stefan Wolf, Jannik Koch, Lars Sommer, and Jürgen Beyerer. Addressing bias in fine-grained classification datasets: A strategy for reliable evaluation. In *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–7, 2023. 2, 4

[47] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 7

[48] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2, 4, 7

[49] Jun Yu, Hao Chang, Keda Lu, Guochen Xie, Liwen Zhang, Zhongpeng Cai, Shenshen Du, Zhihong Wei, Zepeng Liu, Fang Gao, and Feng Shuang. Bag of tricks and a strong baseline for fgvc. In Faggioli et al. [8], pages 2275–2290. 3

[50] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regular-

ization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 5

[51] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 5

[52] Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5531–5540, January 2023. 3

[53] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13130–13137, Apr. 2020. 7