A Baseline for Cross-Domain Fine-Grained Vehicle Classification in a Supervised Partially Zero-Shot Setting

Stefan Wolf

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
stefan.wolf@kit.edu

Abstract

Fine-grained vehicle classification is an important task particularly for security applications like searching for cars of suspects who abuse stolen license plates. However, data privacy and the large number of existing car models render it highly difficult to create a large up-to-date dataset for fine-grained vehicle classification with surveillance images. While a large number of images of vehicles are available in the web due to car selling sites, they have a perspective which is vastly different to surveillance images. Domain adaptation is the field of research that uses domain-wise inappropriate images for training of classification models with the target of running accurate inference on images of a different domain. Since the widely considered unsupervised and semi-supervised domain adaptation settings are unrealistic for fine-grained vehicle classification, we establish a baseline for cross-domain fine-grained vehicle classification in a supervised partially zero-shot setting. Our results indicate that existing domain adaptation methods like domain adversarial training and triplet loss are still advantageous for this setting and we show the benefit of distance-based classification for this task.

1 Introduction

Fine-grained classification tasks like vehicle make and model recognition are relying on large datasets for training. These are needed since the small inter-class variance compared to the large intra-class variance are required to be properly approximated by the learned model. While in the web, a large amount of images for different cars are provided by e.g. car selling sites, fine-grained classification is often applied in different domains. For example, vehicle make and model recognition is useful for security applications like manhunt when applied to cameras on highways which provide a surveillance perspective. However, for these perspectives, the availability of data is scarce. The situation is worsened by the high rate of car manufacturers proposing new vehicle models .

To approach the lack of data, domain adaptation methods can enable the use of the large-scale availability of data of different domains like web-nature images to perform tasks like classification in domains which have a limited availability of data like surveillance. While domain adaptation has been widely approached [33, 45, 48, 22, 4, 25, 11] and also specifically for fine-grained classification [10, 31, 34, 35, 44] applications, an unsupervised or semi-supervised domain adaptation setting is commonly assumed. In these settings, a large number of images is present in the target domain for all classes but the labels aren't present for any or only a part of the images. However, for real-world use-cases, the assumption to have data for all classes is hard to fulfill since it can only be assured if labels would be present. Thus, we focus on a different domain adaptation setting: a supervised partially zero-shot setting [38]. This setting assumes that for a large number of base classes, images and labels are available for both domains while for a small number of novel classes, images and labels are only available for the source domain. For these novel classes, no images are available at all for the target domain during training. However, the evaluation on these novel classes with images from the target domain is the main focus of the setting.

Since the research for such a setting is rather small [32], we provide an extensive evaluation of existing domain adaptation methods to find a good baseline for further research. Besides the widely applied domain adversarial learning [8], we explore the use of metric learning with a triplet loss which also has shown advantages for classification across domains [20, 32].

Based on these experiments, we found that a typical softmax classifier only achieves a low classification accuracy for the novel classes. However, a domain adversarial loss heavily increases the accuracy. A distance-based classifier with a combination of a cross entropy loss and a triplet loss showed promising results which can further be improved by the use of a domain adversarial loss resulting in the overall best model.

In Section 2, existing works in the fields of fine-grained classification, cross-domain classification and cross-domain fine-grained classification are introduced. In Section 3.1, the evaluated methods are described and the evaluation results are shown in Section 4. A conclusion of this work is given in Section 4.

2 Related work

In this chapter, an overview of the literature in the fields of fine-grained classification and cross-domain classification as well as works which employ cross-domain classification for fine-grained classification tasks is given.

2.1 Fine-grained classification

Various approaches have been used to improve the accuracy for fine-grained classification. While all recent approaches share their basis of deep neural networks, there are several different extensions and they can be structured into the following categories. **Part-based models** first detect relevant regions like specific parts of a vehicle before the crops of these parts are fed into a convolutional neural network (CNN) [7, 15, 28, 41]. This reduces the feature space to significant parts and thus, reduces the risk of overfitting. **Bilinear CNNs** employ two networks to separate the localization and the extraction of important features. The networks are combined by calculating the outer product of both resulting feature vectors [23]. Several extensions have been proposed to improve the accuracy and efficiency of bilinear CNNs [9, 19, 43]. Multiple authors employ **multi-task learning** by learning an auxiliary task like predicting the viewpoint of the image that provides support for the main task of fine-grained classification. The auxiliary task is performed only during

training to improve the learned features [3] or also during inference to provide the network with additional information [24, 29]. **Hierarchical classification** exploits that fine-grained categories are usually defined on multiple layers, e.g. make, model and year of a car. This technique was explored by training multiple layers of the hierarchy in a round-robin manner [16] and by training cascaded classifiers [2]. **Metric learning** has also been applied to improve the features by minimizing intra-class variance and maximizing inter-class variance [17, 30, 42]. **Temporal classification** uses videos as input modality for fine-grained object recognition [1, 46, 18] instead of single images as done by most works. **Webly-supervised classification** gathers additional data from the web with image databases like Flickr providing images with additional meta information that can be used for defining labels [6, 39].

2.2 Cross-domain classification

Domain adaptation is usually employed if classification has to be done in a domain for which a lack of data exists. The lack of data can be in the form of missing images or missing annotations. Mostly an unsupervised scenario is considered which contains abundant but unlabaled data for the target domain. To approach a cross-domain setting, multiple methods have been proposed. We follow the taxonomy of Wang and Deng [33] for the categorization of the approaches. **Discrepancy-based domain adaptation** methods are based on a criterion during fine-tuning to increase the accuracy for the target domain. Proposed criteria are class-based [31, 45], statistic-based [48], architecture-based [22] or geometry-based [4]. **Adversarial-based domain adaptation** methods target a domain confusion of the trained network which disables the possibility of exploiting the domain of an image for the classification decision. This can be done by generative approaches which transform the appearance of a source sample such that it can not be distinguished from the distribution of target samples [25]. Non-generative approaches have also been explored by using domain adversarial training with a domain classifier that is preceded by a gradient reversal layer during training. This leads to features which are invariant in regard to distinguishing the domains. **Reconstruction-based domain adaptation** methods reconstruct samples from either domain to the other domain to create a

domain-invariant representation. This has been explored by using a combination of an encoder and a decoder [11] as well as using a Cycle-GAN [47] that keeps semantic information intact by using a cycle-consistency constraint [14].

2.3 Cross-domain fine-grained classification

Some researchers have already addressed fine-grained classification in a cross-domain setting. Gebru et al. [10] exploit the hierarchical nature of fine-grained classification by adding an attribute consistency loss that enforces a matching of coarse-grained attributes like vehicle types to the fine-grained category. With the coarse-grained attribute prediction being a significantly easier task, it is more domain invariant and thus, supports stabilizing the fine-grained prediction due to the new consistency loss. Tzeng et al. [31] and Wang et al. [34] also exploit the attribute and coarse-grained labels inherent to fine-grained classification tasks to improve the domain adaptation. Wang et al. [35] extends adversarial domain-level adaptation by a category-level domain alignment for semi-supervised domain adaptation. Additionally, a part-wise classification to optimize the fine-grained classification accuracy is introduced. Yu et al. [44] achieve a class confusion by training separate class labels for each domain in a pre-training phase and swapping the class labels in a fine-tuning phase with the target of achieving domain confusion while compared to domain adversarial training, keeping the class-separability of the features intact during the adaption process.

3 Methods

In this section, the evaluated methods are described. They can be mainly divided by the type of classification. We evaluate a softmax classifier and a distance-based classifier. As feature extracting backbone, we use ResNet-50 [12] for both variants. On top of both variants, we evaluate the usage of domain adversarial training [8] to improve the domain invariance. While only common for distance-based classification, we evaluate a triplet loss [36] for both variants due to the reported advantages in regards to cross-domain classification [21].

3.1 Softmax classifier

The softmax classifier employs a fully-connected layer to predict as many logits as number of classes and afterwards applies a softmax activation layer to normalize the scores. On top of this output, a cross entropy loss is used to calculate an error measurement.

Additionally, we evaluate the use of a domain adversarial head and an auxiliary triplet loss to improve the domain invariance of the features. Both additions are applied directly on the features of the backbone.

3.2 Distance-based classifier

For the distance-based classifier, during inference, we feed each preprocessed image into the backbone network and calculate the distance between the feature vector of the sample and a prototype feature vector for each class. We choose the class as final prediction for which the distance has the lowest value. The prototype is calculated as the mean of all training samples of a class from the source domain. We also evaluated the use of a medoid instead of a mean but the results indicated an advantage for the mean. Regarding the distance measure, we evaluated the euclidean norm and the negative cosine similarity with the results showing a clear advantage for the negative cosine similarity while the euclidean norm usually prevented the network from converging properly. Since the cosine similarity is originally a similarity instead of a distance measure, we use the negative of the cosine similarity as distance measure. The classification can be described by the following formulas:

$$\mathbf{p_c} = \frac{1}{|X_c|} \sum_{x \in X_c} f(x) \tag{3.1}$$

$$c(x) = \underset{c \in C}{\operatorname{argmin}} - \frac{f(x) \cdot \mathbf{p_c}}{\|f(x)\| \|\mathbf{p_c}\|} \tag{3.2}$$

where $\mathbf{p_c}$ is the feature prototype for the class c, X_c is the set of training images of a class c from the source domain, f is the backbone feature extractor, $c(x)$ is the predicted class for an image x and C is the set of known classes.

During training, we apply a cross entropy loss with a softmax activation on top of a fully-connected layer. Since the cross entropy loss tends to learn features which are highly dependent on the domain, we use a triplet loss as additional loss function that regularizes the network in regards to the domains. Additionally, the triplet loss ensures that the chosen distance measure is appropriate for the features during inference. After training, the fully-connected layer is dropped and the extracted features are directly used as described above.

3.3 Domain adversarial training

Ganin et al. [8] proposed a domain adversarial training method. It applies a simple domain classifier on top of the features extracted by the backbone and inserts a gradient reversal layer between the network and the domain classifier. The gradient reversal layer leads to learning features which are most inappropriate for a classification of the domain and thus, the features are expected to be invariant in regards to the domain. Therefore, the classification loss which is applied in parallel will focus on learning features which are inherent to the class instead of exploiting the domain.

For the domain classification head, we employ two hidden fully-connected layers with 1024 channels with each being followed by a ReLU activation and a batch normalization layer. A final fully-connected layer with a single output channel which is followed by a sigmoid activation predicts the domain. A binary cross entropy is applied as training loss for the domain classification.

The gradient reversal layer includes a gating that controls the influence of the reversed gradient of the domain classification loss onto the main network. We call this parameter λ. A λ of 1 means an unhindered influence while a λ of 0 means that the domain classification has no influence on the main network at all. A good choice of λ might depend on the current state of training and a pre-set value is probably not appropriate. Our results showed that the loss coupling of λ proposed by Wiedemer et al. [37] was superior to a pre-set value and an increasing schedule of λ as it was originally proposed for the domain adversarial training [8]. The loss coupling sets λ for each iteration based on the domain classification loss value of the previous iteration. The exact formula is $\lambda_i = \exp(-L_{d,i-1})$ with λ_i being the set λ for the iteration i and $L_{d,i-1}$ being

the domain classification loss for iteration $i-1$. This ensures that the domain classification only has a strong influence on the main network if the loss is low meaning that the domain classifier is able to classify the domain adequately. In case of a high domain loss, the domain classifier is not able to classify the domains properly and will not provide a good domain adversarial loss.

3.4 Triplet loss

A triplet loss [36] explicitly minimizes the distance of features of the same class while maximizing the distance of features of different classes with respect to a chosen distance measure. While the cross entropy loss also tends to show a similar behavior, it only enforces a linear separability of classes which can result in features of a single class still being spread in feature space. This can be particularly dramatic for cross-domain scenarios for which the distribution of images is different between training and inference. Thus, we apply a triplet loss as additional loss that directly minimizes the distance of features of the same class.

4 Experiments

We execute quantitative evaluations to find a good baseline for cross-domain classification under a supervised partially zero-shot setting. First, the settings of the comparisons are described. Afterwards, the results are discussed. The comparisons include ablation studies for a softmax classifier, ablation studies for a distance-based classifier and a comparison between both approaches.

4.1 Settings

The datasets used for the experiments are described first. Afterwards, the evaluation metrics and training details are reported.

4.1.1 Dataset

As dataset, we choose CompCars [40] which is one of the largest fine-grained vehicle classification datasets available and consists of a web-nature part (CompCars Web) and a surveillance-nature part (CompCars SV). The CompCars Web has a predefined split of 16.016 training images and 14.939 test images. The predefined split of the CompCars SV contains 31.148 training images and 13.333 test images.

While the CompCars Web is labeled according to the make, model and year of a specific car, the CompCars SV is only labeled up to the model of a car and lacks the year as annotation. Thus, we also only consider the model for all cars in CompCars Web. This results in a total of 431 classes for CompCars Web and a total of 281 classes for CompCars SV. We identify the intersection of both sets of classes and use only these for our experiments. Thus, we consider a total of 181 classes. Based on this set of classes, we create three different random splits of base and novel classes with the base classes containing 90% and the novel classes containing 10% of the classes. While during training, for the base classes abundant labeled images are available in both domains, we restrict the availability of data for the novel classes to the source domain of CompCars Web and no images from CompCars SV are available for the novel classes. For each experiment, a model is trained and evaluated on each split and the results are averaged.

4.1.2 Evaluation metric

We use the F1 score on the CompCars SV as main metric for our experiments. We report the class-wise F1 score averaged over the base and the novel classes separately. Since our focus is on adding new classes to the classification, we focus mainly on the F1 score of the novel classes. Due to images of all classes being included in the test set, base classes still influence the score of the novel classes and vice versa. This is sensible since a network only focused on the prediction of novel classes should still be able to distinguish them from the total of all base classes even when distinguishing the base classes might be of minor importance.

4.1.3 Training details

We choose SGD as optimizer with an initial learning rate of 0.04 and a learning rate reduction by $10\times$ is applied after 2500 iterations. We apply a momentum of 0.9 and a weight decay of 10^{-4}. The training is running for 12000 iterations in total. A batch size of 512 per GPU with two GPUs is used. Each batch contains 256 Web and 256 SV images. We evaluate after every 1000 iterations and apply early-stopping by choosing the checkpoint with the highest F1 score for novel classes on the CompCars SV images. The weights are initialized from a model pre-trained on ImageNet. During training, for each image, a crop spanning an area between 8% and 100% of the original image is taken randomly and is resized to a size of 224×224 pixels afterwards. Additionally, a random horizontal flip is applied with 50% probability. Afterwards, the image is normalized using the mean and the standard deviation values of the pre-training on ImageNet. For experiments with a triplet loss, we employ hard negative mining [13] and a margin of 0.3 since preliminary experiments have shown good results for this value.

4.2 Inference details

During evaluation, the images are resized such that the shorter side has 256 pixels while keeping the aspect ratio. Afterwards, a crop of size 224×224 pixels is taken from the center of the resized image. The normalization is applied similar to the training configuration.

4.3 Softmax classification

We evaluate a softmax classifier as the most common architecture for deep-learning-based classification. Since softmax classifiers tend to heavily exploit domains in the classification, we explore the use of domain adversarial training and an auxiliary triplet loss to improve the domain invariance of the network.

Adversarial training	λ-schedule	λ base value	Base F1	Novel F1
No	-	-	95.4	43.0
Yes	Constant	0.1	96.4	66.3
Yes	Increasing	0.1	96.4	66.3
Yes	Increasing	1.0	96.4	65.9
Yes	Coupled	0.1	**96.5**	**67.7**
Yes	Coupled	1.0	96.0	66.4

Table 4.1: Evaluation of different schedules for the λ parameter of the domain adversarial training. The results indicate a clear advantage for the coupled schedule when focusing on the important novel classes.

4.3.1 Domain adversarial training

Adversarial domain adaptation [8] is a widely applied approach for domain adaptation. In order to find a strong baseline, we evaluate different schedules of the λ parameter that controls the influence of the domain adversarial head onto the main network. Besides a constant value and a widely applied monotonically increasing schedule [8], the coupled schedule by Wiedemer et al. [37] is also evaluated. The set λ base value describes the constant value for the constant schedule, the maximum value for the increasing schedule and the highest possible value (in case of zero domain classification loss) for the coupled schedule. The results are shown in Table 4.1.

The adversarial training leads to a large improvement of the base F1 score but particularly of the novel F1 score with all evaluated schedules for λ. While the impact of the schedule for λ is negligible for the base F1 score, for the important novel F1 score, the best results are achieved with the coupled schedule and a λ base value of 0.1. Based on these results, we continue to use these settings for all further experiments involving adversarial domain adaptation. The adversarial training reduces the impact of the domain onto the features and thus, leads to features of novel classes in the target domain being closer to features of the same class in the source domain. Therefore, the samples of novel classes in the target

Triplet loss	Base F1	Novel F1
No	95.4	43.0
Yes	**95.9**	**54.5**

Table 4.2: Evaluation of a triplet loss as auxiliary loss for a softmax classifier. The results indicate that an auxiliary triplet loss can improve the domain invariance of a softmax classifier.

domain are classified more accurately which in turn leads to less confusion with base classes. Thus, also the base class accuracy is improved.

4.3.2 Auxiliary triplet loss

The triplet loss has shown to be more domain invariant than a pure cross entropy loss. Thus, we evaluate the impact of an auxiliary triplet loss in Table 4.2. The triplet loss uses the negative cosine similarity as distance measure. A training with euclidean norm as distance measure did not converge properly since the euclidean norm enforces a feature space that is not well suited for the cross entropy loss. Thus, results for the euclidean norm are not reported.

The results show a clear advantage of the triplet loss for the accuracy of the base as well as the novel classes. The increase is probably a result of the triplet loss forcing a distance of close to zero in feature space for all samples of a class and thus, reducing the possibility of a spread due to different domains. While this only applies for the base classes in training, it probably also reduces the distance of samples of the novel classes between both domains leading to the improvement in the novel class accuracy. This improvement then leads to an improvement in base class accuracy due to less confusion with novel classes occurring.

Distance measure	Base F1	Novel F1
Euclidean norm	8.4	6.2
Negative cosine similarity	**96.0**	**62.9**

Table 4.3: Comparing distance measures for a distance-based classifier. The negative cosine similarity shows a strong advantage with the euclidean norm showing poor results due to the cross entropy loss not converging properly.

4.4 Distance-based classification

While CNNs are mostly combined with a logit-based classification head, distance-based classification and metric learning provide a higher flexibility due to not limiting the model to a specific set of classes during training.

4.4.1 Distance measure

For the distance-based classification, the choice of the distance measure is a crucial parameter. Thus, we compare the use of an euclidean norm as well as negative cosine similarity. The respective distance measure is applied for the triplet loss as well as for the classification. The results of the comparison are shown in Table 4.3. They indicate a strong advantage of the negative cosine similarity while the training with the euclidean norm does not properly converge. Particularly, the training of the triplet loss with an euclidean norm leads to a non-decreasing cross entropy loss. The embedding induced by a triplet loss with an euclidean norm seems to be incompatible with a logit-based softmax classification and a cross entropy loss. Seemingly, the optimizer can not converge to a proper embedding which suits both losses.

4.4.2 Prototype aggregation

For the classification, we aggregate all training samples from the source domain to estimate a prototype for each class and choose the class whose prototype is the closest to the input samples in terms of feature distance. For the aggregation

Aggregation	Base F1	Novel F1
Mean	**96.0**	**62.9**
Medoid	**96.0**	61.8

Table 4.4: Comparison of the estimation methods for the class prototype. Using the mean of the train samples shows a significant advantage over using the medoid.

Domain adversarial training	Base F1	Novel F1
No	96.0	62.9
Yes	**96.3**	**69.8**

Table 4.5: Evaluation of applying domain adversarial training with distance-based classification. The results show that adversarial training can provide an advantage in combination with a distance-based classifier.

of the samples, we evaluate a mean of the features and a medoid of the features. The medoid is defined as the sample which has the smallest total distance to all other samples. The results are shown in Table 4.4. While the difference on the base classes is negligible, the mean aggregation shows a clear advantage over the medoid for the novel classes.

4.4.3 Domain adversarial training

While the triplet loss already provides a strong improvement in terms of domain invariance for distance-based classification, we evaluate if domain adversarial training can still lead to an improved accuracy. Therefore, we apply domain adversarial training with the best setting as in the previous ablation studies additional to the cross entropy loss and the triplet loss we commonly use for the distance-based classifier. The results are shown in Table 4.5 and indicate a slight increase in terms of base class accuracy and a high increase in terms of novel class accuracy.

Method	Base F1	Novel F1
Softmax classifier	95.4	43.0
Softmax classifier with adversarial training	**96.5**	67.7
Softmax classifier with triplet loss	95.9	54.5
Distance-based classifier	96.0	62.9
Distance-based classifier with adversarial training	96.3	**69.8**

Table 4.6: Comparison of softmax classifiers with and without domain regularization methods and distance-based classification. The results show the advantage of distance-based classification for the accuracy of the novel classes while the softmax classifier with domain adversarial training shows a slight advantage for the base class accuracy.

4.5 Comparison of softmax classification and distance-based classification

We compare softmax-based classification methods with and without domain adaptation extensions to a distance-based classification method in Table 4.6. For the softmax-based classification, a domain adversarial training as well as an auxiliary triplet loss is evaluated to improve cross-domain classification accuracy.

While the softmax classifier with the adversarial training shows the highest accuracy for the base classes, the distance-based classifier combined with a domain adversarial training follows closely behind and has a significant advantage in terms of novel class accuracy compared to all evaluated distance-based classifiers. Without adversarial training, the softmax classifier shows a heavy drop in accuracy particularly of the novel classes. The triplet loss also provides a large benefit for the softmax classifier. However, it still shows a large accuracy gap when compared to the adversarial loss.

5 Conclusion

In this work, different domain adaptation approaches were evaluated in a supervised partially zero-shot setting for fine-grained vehicle classification to employ web images as training data for classification on surveillance images. The results show the importance of domain adversarial training to achieve acceptable results with a softmax-based classifier. However, a distance-based classifier employing a combination of a cross entropy loss and a triplet loss still show competitive results which can still be improved by domain adversarial training. This combination showed the overall best results for the classification of the novel classes in our evaluation.

Evaluation of better backbones as modern vision transformers [26, 5] or state-of-the-art convolutional network architectures [27] is up to future work. Other areas of future research are improvements directly targeting the supervised partially zero-shot setting which have not yet been evaluated for other settings.

References

[1] Yousef Alsahafi et al. “CarVideos: A Novel Dataset for Fine-Grained Car Classification in Videos”. In: *16th International Conference on Information Technology-New Generations (ITNG 2019)*. 2019.

[2] Marco Buzzelli and Luca Segantin. “Revisiting the CompCars Dataset for Hierarchical Car Classification: New Annotations, Experiments, and Results”. In: *Sensors* 21.2 (2021).

[3] Qianqiu Chen, Wei Liu, and Xiaoxia Yu. “A Viewpoint Aware Multi-Task Learning Framework for Fine-Grained Vehicle Recognition”. In: *IEEE Access* 8 (2020), pp. 171912–171923.

[4] Sumit Chopra, Suhrid Balakrishnan, and Raghuraman Gopalan. “Dlid: Deep learning for domain adaptation by interpolating between domains”. In: *ICML workshop on challenges in representation learning*. Vol. 2. 6. 2013.

[5] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations*. 2021.

[6] Haodong Duan et al. “Omni-Sourced Webly-Supervised Learning for Video Recognition”. In: *Computer Vision – ECCV 2020*. 2020.

[7] Jie Fang et al. “Fine-Grained Vehicle Model Recognition Using A Coarse-to-Fine Convolutional Neural Network Architecture”. In: *IEEE Transactions on Intelligent Transportation Systems* 18.7 (2017), pp. 1782–1792.

[8] Yaroslav Ganin et al. “Domain-Adversarial Training of Neural Networks”. In: *Journal of Machine Learning Research* 17.59 (2016), pp. 1–35.

[9] Yang Gao et al. “Compact Bilinear Pooling”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[10] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. “Fine-Grained Recognition in the Wild: A Multi-Task Domain Adaptation Approach”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[11] Muhammad Ghifary et al. “Domain Generalization for Object Recognition With Multi-Task Autoencoders”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[12] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. “In defense of the triplet loss for person re-identification”. In: *arXiv preprint arXiv:1703.07737* (2017).

[14] Judy Hoffman et al. “CyCADA: Cycle-Consistent Adversarial Domain Adaptation”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1989–1998.

[15] Shaoli Huang et al. “Part-Stacked CNN for Fine-Grained Visual Categorization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[16] Yuqi Huo et al. “Coarse-to-Fine Grained Classification”. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR’19. 2019.

[17] Alper Kayabasi, Kaan Karaman, and Ibrahim Batuhan Akkaya. “Comparison of distance metric learning methods against label noise for fine-grained recognition”. In: *Automatic Target Recognition XXXI*. Vol. 11729. 2021.

[18] Jannik Koch, Stefan Wolf, and Jürgen Beyerer. “A Transformer-Based Late-Fusion Mechanism for Fine-Grained Object Recognition in Videos”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2023.

[19] Shu Kong and Charless Fowlkes. “Low-Rank Bilinear Pooling for Fine-Grained Classification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[20] Pablo Laiz, Jordi Vitria, and Santi Segui. “Using the Triplet Loss for Domain Adaptation in WCE”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2019.

[21] Pablo Laiz, Jordi Vitria, and Santi Segui. “Using the Triplet Loss for Domain Adaptation in WCE”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2019.

[22] Yanghao Li et al. “Adaptive Batch Normalization for practical domain adaptation”. In: *Pattern Recognition* 80 (2018), pp. 109–117.

[23] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. “Bilinear CNN Models for Fine-Grained Visual Recognition”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[24] Yen-Liang Lin et al. “Jointly Optimizing 3D Model Fitting and Fine-Grained Classification”. In: *Computer Vision – ECCV 2014*. 2014.

[25] Ming-Yu Liu and Oncel Tuzel. “Coupled Generative Adversarial Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016.

[26] Ze Liu et al. “Swin Transformer V2: Scaling Up Capacity and Resolution”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 12009–12019.

[27] Zhuang Liu et al. “A ConvNet for the 2020s”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 11976–11986.

[28] Marcel Simon and Erik Rodner. “Neural Activation Constellations: Unsupervised Part Model Discovery With Convolutional Networks”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[29] Jakub Sochor, Adam Herout, and Jiri Havel. “BoxCars: 3D Boxes as CNN Input for Improved Fine-Grained Vehicle Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[30] Kihyuk Sohn. “Improved Deep Metric Learning with Multi-class N-pair Loss Objective”. In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016.

[31] Eric Tzeng et al. “Simultaneous Deep Transfer Across Domains and Tasks”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[32] Naoto Usuyama et al. “ePillID Dataset: A Low-Shot Fine-Grained Benchmark for Pill Identification”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2020.

[33] Mei Wang and Weihong Deng. “Deep visual domain adaptation: A survey”. In: *Neurocomputing* 312 (2018), pp. 135–153.

[34] Sinan Wang et al. “Progressive Adversarial Networks for Fine-Grained Domain Adaptation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[35] Yimu Wang et al. “An Adversarial Domain Adaptation Network for Cross-Domain Fine-Grained Recognition”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2020.

[36] Kilian Q. Weinberger and Lawrence K. Saul. “Distance Metric Learning for Large Margin Nearest Neighbor Classification”. In: *Journal of Machine Learning Research* 10.9 (2009), pp. 207–244.

[37] Thaddäus Wiedemer et al. “Few-Shot Supervised Prototype Alignment for Pedestrian Detection on Fisheye Images”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2022, pp. 4142–4153.

[38] Stefan Wolf. “Cross-Domain Fine-Grained Classification: A Review”. In: *Proceedings of the 2021 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory*. Vol. 54. Karlsruher Institut für Technologie (KIT), 2022, pp. 189–205. ISBN: 978-3-7315-1171-7.

[39] Zhe Xu et al. “Webly-Supervised Fine-Grained Visual Categorization via Deep Domain Adaptation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.5 (2018), pp. 1100–1113.

[40] Linjie Yang et al. “A Large-Scale Car Dataset for Fine-Grained Categorization and Verification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[41] Hantao Yao et al. “Coarse-to-Fine Description for Fine-Grained Visual Categorization”. In: *IEEE Transactions on Image Processing* 25.10 (2016), pp. 4858–4872.

[42] Baosheng Yu et al. “Correcting the Triplet Selection Bias for Triplet Loss”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[43] Chaojian Yu et al. “Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018.

[44] Han Yu, Rong Jiang, and Aiping Li. “Striking a Balance in Unsupervised Fine-Grained Domain Adaptation Using Adversarial Learning”. In: *Knowledge Science, Engineering and Management*. 2020.

[45] Xu Zhang et al. *Deep Transfer Network: Unsupervised Domain Adaptation*. arXiv:1503.00591 [cs.CV]. 2015.

[46] Chen Zhu et al. “Fine-grained Video Categorization with Redundancy Reduction Attention”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[47] Jun-Yan Zhu et al. “Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[48] Fuzhen Zhuang et al. “Supervised Representation Learning: Transfer Learning with Deep Autoencoders”. In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI’15. 2015.