Supervised Domain Adaptation with Disjoint Label Spaces for Fine-Grained Classification

Enrico Krohmer[1,2,3], Stefan Wolf[2,1,3], and Jürgen Beyerer[1,2,3]

[1] Fraunhofer IOSB, Institute of Optronics, System Technologies and Image Exploitation, Fraunhoferstrasse 1, 76131 Karlsruhe, Germany
[2] Vision and Fusion Lab, Karlsruhe Institute of Technology KIT, Vincenz-Prießnitz-Straße 3, 76131 Karlsruhe, Germany
[3] Fraunhofer Center for Machine Learning

Abstract. Domain adaptation scenarios commonly assume that the label spaces of the source and target domains are either equal or share a common set of classes. However, in fine-grained classification settings, it is likely that the common label set is empty. Therefore, we approach a supervised domain adaptation scenario where the label spaces of the source and target domains are available but disjoint during training. The classifier is tasked with generalizing to the complete target domain where classes are not only from the target label space but also from the source label space. We introduce a novel CycleGAN variant, FCC-GAN, which translates source images into target-stylized images that preserve their class-specific features. To further encourage the classifier to learn domain-invariant representations, we pre-train the classifier exclusively on the target domain and then employ supervised contrastive learning on source, target, and target-stylized images. We demonstrate that this framework outperforms existing domain adaptation methods in a fine-grained classification task under the disjoint label space assumption. Code and supplementary material is available at: https://github.com/enricokrohmer/sda_dls.

Keywords: Fine-Grained Classification · Supervised Domain Adaptation · Disjoint Label Spaces

1 Introduction

Fine-grained classification tasks like vehicle make and model recognition have been dominated by data-intensive deep learning methods recently. While good results have been achieved under the assumption of enough data available, the acquisition of images and image labels are an enormous task for fine-grained classification. Particularly, fine-grained classification requires large amount of training data due to the specific class-distinguishing features often being difficult to extract. Additionally, the high specificity of the classes renders it challenging to find a large amount of appropriate samples for each class. For this reason, domain adaptation methods have been investigated for fine-grained classification

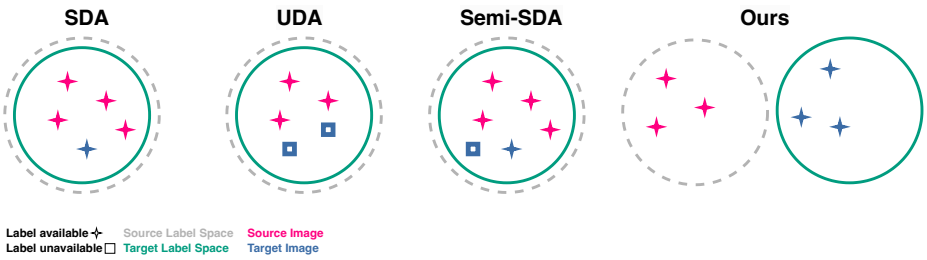

Fig. 1: Visualization of different domain adaptation scenarios. In our domain adaptation framework, we assume that the source and target domains have disjoint label spaces, contrasting with other approaches that assume identical label spaces. This means each domain contains images belonging only to its specific set of classes.

which exploit domains with large amounts of well-annotated samples available as a source domain which provides information to learn class-distinguishing feature. Typical examples of source domains for fine-grained domain adaptation are web images of vehicles from car-selling sites or field guides for the classification of birds [9,37,38,42]. In domain adaptation scenarios, the final classification is performed in a target domain different from the source domain. While the source domain has abundant annotated samples available, the target domain shows some form of limited availability of data. However, the existing fine-grained domain adaptation studies only explore unsupervised or semi-supervised domain adaptation scenarios which consider a common label space for the source and target domain. This includes the presence of samples for all classes in the target domain, a hard to fulfill claim for fine-grained classification with a high number of classes, particularly, assuming the lack of labels in unsupervised domain adaptation.

Thus, we approach a different domain adaptation scenario with disjoint label spaces in the source and target domain, not requiring the same classes to be present in the source and target domain. Nonetheless, the final classifier is tasked to perform classification on target domain images for the join of the source and the target label space. So, it has to support all classes. To still be a viable scenario, we consider the presence of labels in the target domain necessary and thus, it is a supervised domain adaptation scenario. This enables a more flexible usage of domain adaptation. We can exploit a supervised dataset in the target domain, which is already commonly available for most relevant domains, and additional classes can be added to the classification solely based on samples from the source domain.

For example, for fine-grained vehicle classification in surveillance scenarios, data privacy limits the possibility to acquire data in public space drastically. However, the constant introduction of new vehicle models by the manufacturers requires regular update of training datasets and thus, acquiring new data. Therefore, we aim to exploit the limited available data best possible. For the target domain of surveillance images, annotated datasets are publicly available [32,39].

So, the main challenge is to extend it with new vehicle models which we solve with a disjoint label spaces domain adaptation scenario.

Banatt *et al*. [1] explored a similar scenario and showed that the disjoint label space leads to significant negative transfer. Thus, the classification performance on the source classes drops below random guessing while also existing domain adaptation methods not proving to be effective in this setting. However, the authors also showed that the negative transfer can be reduced when a small proportion of target domain samples of the source classes are added to the training. We show that a similar effect can be achieved when target-stylized source samples are used to approximate the target domain distribution of the source classes. We approach the scenario with a novel domain adaptation framework which we essay in this study for a fine-grained vehicle classification task with synthetic images as source domain and real surveillance images as target domain. First on, we perform a pre-training on the target domain only including the target label space which focuses on the extraction of domain-specific task-important features. Afterwards, we extend the classification label space by the source domain, in our case, a synthetic dataset of surveillance images of vehicles [31]. Samples of the target domains for the new classes are not required, neither unsupervised nor supervised. Our novel method enhances the CycleGAN [46] to tackle the challenges of applying it for fine-grained domain adaption. While image-to-image translation is a common pattern for domain adaptation, particularly synthetic-to-real adaptation, it has yet to be investigated for fine-grained domain adaptation and its specific challenges. CycleGAN is well capable of generating images of the target domain with the meta-class being well reconstructed. Nonetheless, it lacks the ability to reconstruct fine-grained details such as the logo of the vehicle make. Thus, we propose a feature consistency loss which imposes recovering fine-grained details important for classification. With our feature-consistent CycleGAN, we can transform images from the synthetic source domain to the real target domain while preserving fine details in the car. Based on a model pretrained only on the target images, we use the target-stylized source images as training data combined with the real target images. Thus, we have target images available for all classes, for some classes actual real target images and for some classes pseudo target images which have been target-stylized from source images. We confirm the impact of this enhancement quantitatively and qualitatively. Additional to the classification loss, we apply a supervised contrastive loss [15] to diminish remaining feature information loss between the original source images and the target-stylized counterpart.

Our contributions can be summarized as follows:

1. We are the first to explore domain adaptation methods in depth for a domain adaptation scenario with disjoint label spaces, indicating the deficiencies of the existing methods for this scenario.
2. We propose a novel feature-consistency loss enhancing CycleGAN [46] by improving the capability to distinguish task-relevant details in the image which need to be transferred and domain-relevant details that should be adapted to the new domain.

3. We propose the integration of the supervised contrastive loss [15] for the classification training to diminish the remaining feature information loss between target-stylized source images and their original counterparts.

2 Related Work

Domain Adaptation Scenarios. Multiple domain adaptation scenarios have been proposed to address varying assumptions about label spaces and availability of labels as data availability can differ by application. Regarding assumptions about the label space, these scenarios include a closed-set, multiple open-set, a partial and a universal domain adaptation scenario. In the traditional closed-set scenario, the label spaces of the source and the target domains are assumed to be identical. However, due to applications of domain adaptation are commonly data-limited, especially concerning the presence of classes in datasets, more flexible scenarios were proposed and investigated. Busto and Gall [4] propose the open-set domain adaptation scenario assuming the existence of a set of common classes and additionally samples with an unspecified class for the source and the target domain. However, samples without a specified class are not required to be recognized by their class but recognizing them as an unknown class sample is enough. Saito *et al*. [28] propose a more stringent variant by removing the availability of unknown class samples in the source domain during training. Still, samples of unknown classes in the target domain have to be distinguished from known classes. Cao *et al*. [5] and Zhang *et al*. [44] introduce the partial domain adaptation scenario, where the target label space is a subset of the source label space. Universal domain adaptation [41] includes private classes in the source and the target domain beside a set of common classes. However, all of these scenarios consider private classes as not to be distinguished. Additionally, none of these scenarios consider completely separate label spaces. Nonetheless, both conditions are common in a fine-grained domain adaptation scenario. Thus, we consider disjoint label spaces with the join of the source and target classes to be distinguished in the target domain.

Domain adaptation has been approached in unsupervised, semi-supervised and supervised settings. In unsupervised domain adaptation, a large amount of unlabeled images of the target domain is leveraged for the adaptation process [2, 8, 18, 35]. Semi-supervised domain adaptation introduces labels for some of the images of the target domain [8, 40]. In supervised domain adaptation, all samples from the target domain are labeled [13, 16, 25, 26, 34]. To preserve a challenging aspect in supervised domain adaptation scenarios, only a few samples per class in the target domain are used during training. In contrast, in the supervised domain adaptation scenario we consider, the challenge arises not from a scarcity of samples but from a disjoint label space with a complete lack of images for the source classes in the target domain.

Domain Adaptation Methods. Following the categorization of Wang and Deng [36], there are three general categories of domain adaptation methods.

Discrepancy-based methods perform a fine-tuning on the target data to minimize the shift between the domains based on a criterion. The criterion can be a class label, either unsupervised with a pseudo label [45] or supervised [34], a statistic criterion like maximum mean discrepancy [10, 20] or Kullback-Leibler divergence [43] or an architecture criterion like an adaptive batch normalization [17]. Adversarial-based approaches aim to induce domain confusion by either using generative methods [3] which transform images from source domain to target domain or non-generative methods [8]. Reconstruction-based approaches expect a reconstruction process to generate features which are invariant to domain differences while maintaining class-discriminative properties. They can be based on an encoder-decoder combination [11] or an adversarial reconstruction [46]. Our approach can be categorized as a hybrid approach of a generative adversarial-based approach and a discrepancy-based approach based on a class criterion. While semantic consistency losses have been proposed to retain task-important details for generative domain adaptation approaches [3, 14], as we show in our experiments, they still lack the capability of retaining details on a level necessary for fine-grained classification. Thus, we introduce a feature consistency loss.

In the field of fine-grained domain adaptation, methods are based on the approaches for regular domain adaptation and adjusted for fine-grained classification tasks. A common pattern is to exploit coarse-grained labels to improve the domain alignment process in the expectation that the coarse-grained labels can be recognized more consistently across domains [9, 37]. Wang *et al.* [38] propose the integration of a spatial self-attention module to extract more relevant features for the fine-grained classification. Yu *et al.* [42] propose a new adversarial domain adaptation approach based on label switching which better retains fine-grained features. However, in the field of fine-grained domain adaptation, generative adversarial domain adaptation methods are yet to be investigated. To the best of our knowledge, we are the first to explore generative adversarial domain adaptation approaches for fine-grained domain adaptation.

3 Supervised Domain Adaptation with Disjoint Label Spaces

3.1 Problem Setting

We are given a labeled source domain $\mathcal{S}$ and a labeled target domain $\mathcal{T}_{Full} = \mathcal{T} \cup \mathcal{T}^*$ with label space $\mathcal{Y} = \mathcal{Y}_1 \cup \mathcal{Y}_2$, where $\mathcal{Y}_1 \cap \mathcal{Y}_2 = \emptyset$ and $\mathcal{T} \cap \mathcal{T}^* = \emptyset$. However, $\mathcal{S}$ and $\mathcal{T}^*$ only contain data points whose labels are in $\mathcal{Y}_1$, and $\mathcal{T}$ only contains data points whose labels are in $\mathcal{Y}_2$. For simplicity, we will refer to $\mathcal{Y}_2$ as the label space of $\mathcal{T}$, and $\mathcal{Y}_1$ as the label space of both $\mathcal{S}$ and $\mathcal{T}^*$. This setting was first explored by [1] and a comparison of different domain adaptation scenarios can be seen in Fig. 1.

The goal of supervised domain adaptation with disjoint label spaces (SDA-DLS) is to train a classifier capable to generalize to $\mathcal{T}_{Full}$, while only having access to training data from $\mathcal{S}$ and $\mathcal{T}$.

3.2 Technical Challenges

Models that try to tackle fine-grained classification tasks need to learn good representations of each class to mitigate the low inter-class and high intra-class variance. As in traditional domain adaptation, the model additionally needs to bridge the domain gap between $\mathcal{S}$ and $\mathcal{T}^*$. Otherwise, the classifier may learn domain-specific features from $\mathcal{S}$ which hinders the model's ability to generalize to the target domain.

In SDA-DLS, a classifier is prone to learn domain-specific features not only from $\mathcal{S}$ but also from $\mathcal{T}$ to differentiate between classes from $\mathcal{Y}_1$ and $\mathcal{Y}_2$ during training. If the classifier must classify a sample from $\mathcal{T}^*$ during inference, the domain-specific features of the sample could mislead the classifier into thinking that the sample belongs to $\mathcal{T}$. As a result, the classifier is prone to confusing classes from $\mathcal{Y}_1$ with classes from $\mathcal{Y}_2$. This behavior can be seen as a special case of negative transfer as described by [1].

Addressing the issue of negative transfer within the current domain adaptation frameworks presents its own set of problems: Supervised Domain Adaptation (SDA) operates under the assumption that during training, the source and target domains share the same label space. Multiple SDA methods [25, 26] directly utilize label information on the target domain to derive a domain-invariant representation of samples that belong to the same class but originate from different domains. However, during the training phase of SDA-DLS no class is present in both source and target domain, due to the disjoint label spaces.

An alternative approach involves turning to Unsupervised Domain Adaptation (UDA) methods, which do not depend directly on label information. However, they still assume that the label spaces of the source and target domains are identical, as is typical in classical unsupervised domain adaptation, or that they share a common set of labels, as seen in universal domain adaptation [41].

4 Method

This section will introduce our novel approach for tackling the SDA-DLS scenario. We will incrementally introduce new components to a convolutional neural network that minimizes the standard cross-entropy loss $\mathcal{L}_{Task}$ on both the source and target domain. The framework consists of the following training stages:

1. Train a CycleGAN [46] to translate images from $\mathcal{S}$ to $\mathcal{T}$. Class-specific features are preserved after translation, as CycleGAN also optimizes the novel feature-consistency loss.
2. Pre-train a classifier C on only the target domain.
3. Train C on the source, target and translated source images in conjunction to a supervised contrastive loss [15].

4.1 Target-Only Pretraining

First, the classifier is pre-trained exclusively on the target domain. After pretraining is completed, we freeze the weights of the earlier layers and fine-tune

C on both domains. To accelerate the training process during the second phase, only a small subset of the target domain is utilized. We do not completely exclude target data to prevent C from forgetting the class-specific features of $\mathcal{Y}_2$.

During target-only pre-training (TO-PT), C is optimized to extract features that are specialized for fine-grained classification. Nonetheless, earlier layers are should still be able to extract more general features. By freezing these layers in the subsequent step, we maintain their generality and avoid that C learns domain-specific features from the source domain.

4.2 CycleGAN Recap

Our framework employs a CycleGAN [46] to translate images from the source domain $\mathcal{S}$ to the target domain $\mathcal{T}$ and vice versa. The CycleGAN framework consists of two GANs [12] with generators G and F, and discriminators $D_{\mathcal{S}}$ and $D_{\mathcal{T}}$. G is tasked with transforming images $x \in \mathcal{S}$ to images that adopt the style of images from $\mathcal{T}$. The discriminator $D_{\mathcal{T}}$ is tasked with differentiating between images from $\mathcal{T}$ and the transformed images $G(\mathcal{S})$. Conversely, the generator aims to deceive the discriminator into believing that the transformed images are indeed sampled from $\mathcal{T}$. A generator-discriminator pair optimizes the following adversarial loss function introduced by Goodfellow *et al.* [12]:

$$\mathcal{L}_{gan}(G, D, \mathcal{S}, \mathcal{T}) = \mathbb{E}_{z\sim\mathcal{T}} \log D(z) + \mathbb{E}_{x\sim\mathcal{S}} \log(1 - D(G(x))) \tag{1}$$

The loss function for the opposite direction is defined analogously. Additionally, CycleGAN requires that both generators should be inverse to each other such that $F(G(x)) \approx x$ and $G(F(z)) \approx z$, where $x \in \mathcal{S}, z \in \mathcal{T}$. This is achieved by G and F minimizing the cycle-consistency loss $\mathcal{L}_{cyc}$:

$$\begin{aligned}\mathcal{L}_{cyc}(G, F, \mathcal{S}, \mathcal{T}) = \mathbb{E}_{x\sim\mathcal{S}}\|F(G(x)) - x\|_1 \\ + \mathbb{E}_{z\sim\mathcal{T}}\|G(F(z)) - z\|_1\end{aligned} \tag{2}$$

Combining Eq. (1) and Eq. (2) yields the final loss function of CycleGAN:

$$\begin{aligned}\mathcal{L}_{CycleGAN}(G, F, D_{\mathcal{S}}, D_{\mathcal{T}}, \mathcal{S}, \mathcal{T}) = \mathcal{L}_{gan}(G, D_{\mathcal{T}}, \mathcal{S}, \mathcal{T}) \\ + \mathcal{L}_{gan}(F, D_{\mathcal{S}}, \mathcal{T}, \mathcal{S}) \\ + \lambda_{cyc}\mathcal{L}_{cyc}(G, F, \mathcal{S}, \mathcal{T})\end{aligned} \tag{3}$$

Where λ_{cyc} is a hyperparameter. $\mathcal{L}_{CycleGAN}$ is optimized in the following minimax objective:

$$\min_{G,F} \max_{D_{\mathcal{S}}, D_{\mathcal{T}}} \mathcal{L}_{CycleGAN}(G, F, D_{\mathcal{S}}, D_{\mathcal{T}}, \mathcal{S}, \mathcal{T}) \tag{4}$$

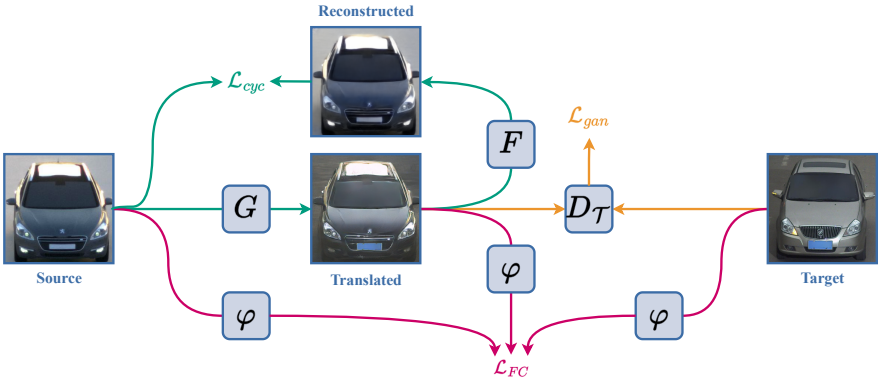

Fig. 2: Translation from $\mathcal{S}$ to $\mathcal{T}$ using FCCGAN. Green arrows show the data flow for the cycle-consistency loss $\mathcal{L}_{cyc}$. Orange arrows show the data flow for the gan loss. Purple arrows show the data flow for the feature-consistency loss. The translation in the opposite direction is performed analogously.

4.3 Feature-Consistency Loss

Translating images via CycleGAN for SDA-DLS has a main caveat. First, due to the disjoint label spaces, the class-specific features in $\mathcal{S}$ are different from the class-specific features in $\mathcal{T}$. As a consequence, the discriminator could easily differentiate between translated and real images if the generator preserves all class-specific features from the input. To minimize $\mathcal{L}_{gan}$, the generators are forced to transform these class-specific features such that they align with the probability distribution of $\mathcal{T}$. In the worst case, label flipping occurs.

To preserve class-specific features after translation and prevent label flipping, we first train a classifier $h \circ \varphi$ by minimizing $\mathcal{L}_{task}$ on the source and target domain. h denotes the classification head of the classifier, and φ denotes its feature extractor. After training, we emit h and freeze the weights of φ. Second, CycleGAN is optimized using $\mathcal{L}_{CycleGAN}$ in conjunction to the novel feature-consistency loss $\mathcal{L}_{FC}$:

$$\mathcal{L}_{FC}(G, \varphi, \mathcal{S}, \mathcal{T}) = \mathbb{E}_{\substack{x \sim \mathcal{S}, \\ z \sim \mathcal{T}}} \Big[\|\varphi(G(x)) - \varphi(x)\|_2 - \|\varphi(G(x)) - \varphi(z)\|_2 + m \Big]_+ \quad (5)$$

At its core, $\mathcal{L}_{FC}$ employs the triplet-loss [29] with margin m where the translated image serves as the anchor, the input image as a positive sample, and an image from the opposite domain as a negative sample. As φ is frozen, the only way to minimize $\mathcal{L}_{FC}$ is by G optimizing $\mathcal{L}_{FC}$ by itself. Therefore, the class-specific features of translated images have to be similar to the features of their original counterpart and distant to class-specific features from $\mathcal{Y}_2$.

We employ the feature-consistency loss for both generators. This leads to the full loss function of our CycleGAN variant, named Feature-Consistent CycleGAN (**FCCGAN**):

$$\begin{aligned}\mathcal{L}_{FCCGAN}(G, F, D_{\mathcal{S}}, D_{\mathcal{T}}, \varphi, \mathcal{S}, \mathcal{T}) = \mathcal{L}_{CycleGAN}(G, F, D_{\mathcal{S}}, D_{\mathcal{T}}, \mathcal{S}, \mathcal{T}) \\ + \lambda_{FC}\mathcal{L}_{FC}(G, \varphi, \mathcal{S}, \mathcal{T}) \\ + \lambda_{FC}\mathcal{L}_{FC}(F, \varphi, \mathcal{T}, \mathcal{S})\end{aligned} \tag{6}$$

Where λ_{FC} is hyperparameter. A visualization of FCCGAN training is shown in Fig. 2.

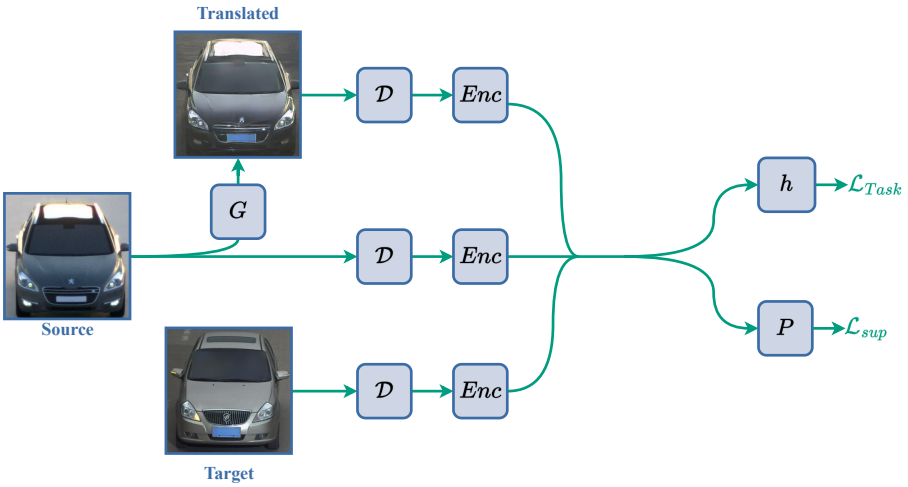

Fig. 3: Visualization of our modified supervised contrastive learning framework. *Enc* denotes the feature extractor of the classifier with classification head h. First, source images are translated via the FCCGAN generator G. Second, the data-augmentation module $\mathcal{D}$ is applied to each image, before being fed to *Enc*. The features are then either fed into h and P to calculate the cross-entropy loss $\mathcal{L}_{Task}$ and the supervised contrastive loss $\mathcal{L}_{sup}$, respectively.

4.4 Feature-Level Alignment

Translated images generated by FCCGAN have no guarantee that they preserve class-specific features perfectly. This implies a trade-off between training with source and translated images: Source images contain all class-specific features but there exists a domain gap between $\mathcal{S}$ and $\mathcal{T}^*$. Images from $G(\mathcal{S})$ only approximate class-specific features but mitigate the domain gap by adapting the style of the target domain. To leverage the advantages of both translated and original data, we present a modification of the supervised contrastive learning framework developed by Khosla *et al.* [15].

Enc denotes the feature extractor of the classifier C. Equivalent to Khosla *et al.* we introduce a data augmentation module $\mathcal{D}$ and a simple feed-forward

Table 1: Comparison between the proposed method and the baselines. All values are rounded to the third decimal.

Model	SSB → CCSV		CCSV → SSB	
	$F_1(\mathcal{Y}_1)$	$F_1(\mathcal{Y}_2)$	$F_1(\mathcal{Y}_1)$	$F_1(\mathcal{Y}_2)$
Direct Transfer	0.660	0.965	0.751	0.978
CycleGAN [46]	0.782	0.968	0.861	0.986
DANN [8]	0.827	0.908	0.629	0.965
DAN [20]	0.895	0.962	0.858	0.985
CyCADA [14]	0.808	0.966	0.800	0.983
Ours	**0.946**	**0.985**	**0.906**	**0.991**
Full Target	0.990	0.971	0.999	0.999

network P. $\mathcal{D}$ applies data augmentations to input images. P projects feature maps produced by Enc into a lower dimensional space. Additionally, the output of P gets normalized.

Given a batch of N source image-label pairs $\{x_i, y_i\}_{i=1}^{N}$ and a batch of target image-label pairs $\{z_i, y_i\}_{i=1}^{N}$. First, we generate $\{G(x_i), y_i\}_{i=1}^{N}$ using the FCC-GAN source-to-target generator. Finally, we calculate the supervised contrastive loss [15] using the concatenated batch of all image-label pairs $\{a_i, y_i\}_{i=1}^{3N}$:

$$\mathcal{L}_{sup} = -\sum_{i=1}^{3N} \frac{1}{|A(i)|} \sum_{j \in A(i)} \log \frac{\exp(p_i \cdot p_j/\tau)}{\sum_{k \neq i} \exp(p_i \cdot p_k/\tau)} \tag{7}$$

Here, $p_i = P(Enc(\mathcal{D}(a_i)))$, $A(i) = \{k \in \mathbb{N} : y_i = y_k, 1 \leq k \leq 3N, k \neq i\}$, τ is a positive temperature parameter, and $\cdot$ denotes the inner product. $\mathcal{L}_{sup}$ forces features within the same class to be close to each other and features from different classes to be distant. Therefore, $\mathcal{L}_{sup}$ encourages Enc to find a domain-invariant representation for source images and their translated counterparts, as they share their label and thus need to be closely aligned in the feature space. Additionally, $\mathcal{L}_{sup}$ aids in addressing the challenges of fine-grained classification tasks: First, it mitigates the low inter-class variance, as images from different classes need to be well separated. Second, it mitigates the high intra-class variance, as samples within the same class need to be close to each other in the feature space.

After target-only pre-training the classifier can minimize $\lambda_{sup}\mathcal{L}_{sup}$ in addition to $\mathcal{L}_{task}$, where λ_{sup} is a hyperparameter. Note that $\mathcal{L}_{task}$ is optimized on source, target and translated source images. A visualization of the proposed method can be seen in Fig. 3.

5 Experiments

5.1 Setup

Datasets. We use the fine-grained vehicle classification datasets **Synset Boulevard** (SSB) [31] and **CompCars Surveillance** (CCSV) [39], specifically employing the Bayer-Bad Configuration of SSB. Both datasets feature surveillance-type vehicle images captured during daytime. There exists a domain gap between SSB and CCSV, as SSB is entirely synthetically generated where as CCSV is a real-world dataset which additionally contains nighttime images. The source and target datasets contain 156 and 281 classes, respectively, with each class denoting a distinct vehicle model. We report results for both adaptation directions, *i.e.* from SSB to CCSV and vice versa. For both directions, we set $\mathcal{Y}_1$ to the 21 classes from SSB and CCSV that match by vehicle model and year. The remaining images in the source domain are omitted, as their classes are not present in the target domain, precluding the possibility of evaluating them. The remaining classes within the respective target domain are designated as $\mathcal{Y}_2$. This leads to 260 classes in $\mathcal{Y}_2$ for SSB to CCSV and 135 classes for CCSV to SSB. Therefore, $\mathcal{S}$ comprises images from the source dataset with classes from $\mathcal{Y}_1$, $\mathcal{T}^*$ comprises images from target dataset with classes from $\mathcal{Y}_1$, and $\mathcal{T}$ comprises the remaining images in the target dataset.

Evaluation Details. We report the F1-Score on $\mathcal{T}_{Full}$ for each method. More specifically, we first compute the class-wise metrics on $\mathcal{T}_{Full}$ and average them based on their associated label spaces. During training, we periodically evaluate each method on a small subset of $\mathcal{T}_{Full}$. This subset serves as our validation set, with the remaining samples comprising the test set. For each method, we report the results computed on the test set for the model instance that achieved the highest F1-Score in the source domain during validation.

Baselines. We report results for a classifier trained on $\mathcal{S}$ and $\mathcal{T}$, a classifier trained on $\mathcal{T}_{Full}$, and a classifier trained on $G(\mathcal{S})$ and $\mathcal{T}$. G denotes the source-to-target generator of a vanilla **CycleGAN** [46]. Additionally, we report results for the following state-of-the-art UDA methods: Domain-Adversarial Neural Networks (**DANN**) [8], Deep Adaptation Networks (**DAN**) [20], and Cycle-Consistent Adversarial Domain Adaptation (**CyCADA**) [14]. We modified each method so that they can be directly applied to the SDA-DLS setting. More specifically, the classifier of each method now minimizes the cross-entropy loss on both the source and target domains, instead of only on the source domain as in the UDA setting.

Implementation. All methods were implemented using the open-source deep learning framework PyTorch [27]. For the classifier and FCCGANs feature extractor we employ a **ConvNeXt-Tiny** [19] that was pre-trained on ImageNet [7]. Features of the classifier are extracted after the final pooling layer. For classifiers

pre-trained on the target domain, we freeze the first three stages. For all other cases, only one stage is frozen. FCCGAN utilizes the generator and discriminator architecture from Zhu *et al.* [46]. We replaced Eq. (1) with a least-square objective [24], and the discriminator is updated based on a history of generated images rather than the most recent one [30]. Both methods were employed by Zhu *et al.* [46] to stabilize CycleGAN training. P is implemented as a feed-forward network with one hidden layer of size 2048 and an output size of 128. We set $\lambda_{cyc} = 10$, $\lambda_{FC} = 10$, the margin of $\mathcal{L}_{FC}$ to 0.5, $\lambda_{sup} = 0.5$, and $\tau = 0.1$.

Training. Each model is trained for 60 epochs using AdamW [22], with betas set to 0.9 and 0.999, and an initial learning rate of 0.0002. We employ cosine annealing [21] without restarts for learning rate scheduling. The weight decay is set to 0.05 for FCCGAN training and to 0.1 for classifier training. For FCCGAN, we set the batch size to 4 images per domain. Classifiers are trained with a batch size of 32. For FCCGAN, we resize each image to 276x276 pixels, then randomly crop them to a size of 256x256, and apply a random flip to each image with a probability of 0.5. For classifier training, we resize each image to 256x256, randomly crop them to 224x224, and apply RandAugment [6]. If a classifier is trained with translated images, the data augmentations are applied post-translation. Within our supervised contrastive learning framework, RandAugment functions as the data-augmentation module $\mathcal{D}$. Additionally, for classifier training, we employ label smoothing [33] with a smoothing parameter of 0.1. We use the same hyperparameters for both directions.

5.2 Results

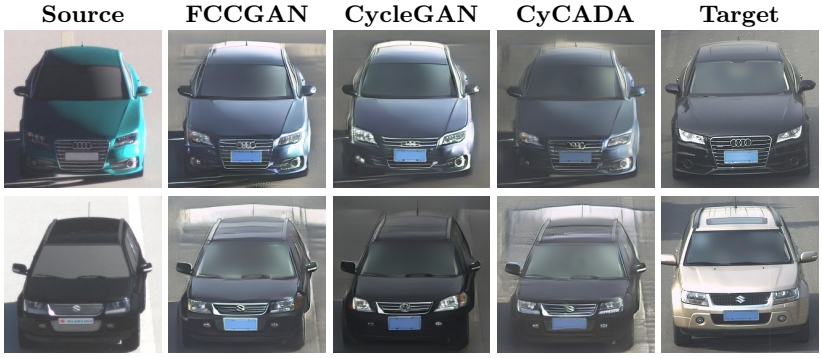

Fig. 4: Images from the source, target and translated source domain. Images in each row belong to the same class.

Classification Results. The classification results are shown in Tab. 1. The Full Target model serves as an upper bound as it represents the optimal sce-

Table 2: Ablation study on SSB to CCSV. All values rounded to the third decimal.

TO-PT	**CycleGAN**	**FCCGAN**	**SupCon**	$F_1(\mathcal{Y}_1)$	$F_1(\mathcal{Y}_2)$
-	-	-	-	0.660	0.965
✓	-	-	-	0.915	0.983
✓	✓	-	-	0.816	0.980
✓	-	✓	-	0.929	0.984
✓	-	✓	✓	**0.946**	**0.985**

nario where training data for all of $\mathcal{T}_{Full}$ is available. Therefore, it achieves the highest F1-Score for classes in $\mathcal{Y}_1$ for both directions. Regardless of the direction, the direct transfer model significantly underperforms on $\mathcal{Y}_1$. A minor drop in performance is observed on $\mathcal{Y}_2$. This observation aligns with the findings of [1].

For classes in $\mathcal{Y}_2$, our method achieves the highest F1-Score in both directions. On SSB to CCSV, our classifier is even able to outperform the classifier trained on the full target domain. This could be attributed to reduced overfitting, as the initial stages are frozen after TO-PT. The UDA methods and the direct transfer approach maintain a relatively high F1-Score for classes from $\mathcal{Y}_2$, as there exists no domain gap between classes from $\mathcal{Y}_2$ during training and inference.

DANN, however, experiences a significant performance decline on $\mathcal{Y}_2$ in both directions. On CCSV to SSB, DANN additionally underperforms to the direct transfer model on $\mathcal{Y}_1$. This indicates that domain adversarial training is not well-suited for the SDA-DLS setting.

Our method significantly outperforms all other domain adaptation methods by combining feature-consistent image translation with domain-invariant feature learning.

DAN achieves the second highest F1-Score on $\mathcal{Y}_1$ for SSB to CCSV and third highest for CCSV to SSB. This suggests that discrepancy-based approaches, such as our method and DAN, are viable strategies for tackling SDA-DLS.

Even though CyCADA employs a semantic consistency loss to preserve features after translation, it fails to match the performance of CycleGAN on CCSV to SSB and only marginally improves over CycleGAN on the opposite direction. This could be due to the second phase of CyCADA training, which employs domain adversarial training. We refer to Fig. 4 for a comparison of image translations for CycleGAN, CyCADA, and FCCGAN.

Ablation Study. We systematically evaluate each component introduced by our framework. The results are shown in Tab. 2. The introduction of TO-PT significantly improves the classifier's ability to generalize to $\mathcal{T}_{Full}$, as it prevents the classifier from learning domain-specific features. However, if the classifier is additionally trained with source images translated by a vanilla CycleGAN, this improvement is mitigated, as class-specific features are lost after translation. Consequently, the classifier is not able to learn a representation that generalizes to $\mathcal{T}^*$. FCCGAN manages to close the domain gap while preserving class-specific

features, thanks to the novel feature-consistency loss. Therefore, the combination of FCCGAN and TO-PT yields an improvement over using TO-PT alone. Nonetheless, the best results are achieved by the complete framework that employs supervised contrastive learning. The classifier learns a domain-invariant representation of source classes, as source and target-stylized images are forced to reside near each other in the feature space. Additionally, the low inter-class variance of fine-grained datasets is mitigated, as samples from distinct classes are forced to be distant from each other.

6 Conclusion

6.1 Summary

In this paper, we approach the supervised domain adaptation scenario with disjoint label spaces for fine-grained classification. We propose FCCGAN, a variant of CycleGAN, designed to translate images in a way that adapts the style of the target domain while preserving class-specific features. To further encourage the classifier to learn domain-invariant representations, we employ supervised contrastive learning and target-only pre-training. Target-only pre-training prevents the classifier from learning domain-specific information from the source domain. Supervised contrastive learning ensures that images from the source and target domains, as well as translated source images, are positioned close to one another in the feature space if they belong to the same class and distanced if not. Operating under the disjoint label space assumption, we evaluate our proposed framework on a fine-grained classification task and demonstrate that it significantly outperforms existing domain adaptation methods.

6.2 Limitations & Future Work

This paper explored four state-of-the-art unsupervised domain adaptation approaches for SDA-DLS. We further encourage the investigation of existing domain adaptation approaches, because UDA methods like DAN [20] are still able to perform relatively well under these new challenges.

Our method was only tested on a fine-grained vehicle classification task. We expect our method to perform well on other datasets *e.g.* birds [37] and aircrafts [23]. However, benchmarking requires fine-grained datasets with a large enough shared label space. Additionally, the domain gap should not involve significant perspective changes, as these can lead to disparities in class-specific features across domains. For example, evaluating on CompCars Web [39] was not feasible because SSB and CCSV include only frontal vehicle images, whereas CompCars Web features vehicles from various perspectives, such as the side and back. This mismatch would mean that for the same vehicle model class-specific features in the source domain might not cover all class-specific features in the target domain. Therefore, the curation of new fine-grained datasets is needed to further generate insights for the supervised domain adaptation with disjoint label spaces scenario.

References

1. Banatt, E., Rajendran, V., Packer, L.: Target domain data induces negative transfer in mixed domain training with disjoint classes. arXiv preprint arXiv:2303.01003 (2023) 3, 5, 6, 13
2. Boqing Gong, Yuan Shi, Fei Sha, Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: CVPR. pp. 2066–2073. IEEE, Providence, RI (Jun 2012). https://doi.org/10.1109/CVPR.2012.6247911 4
3. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks. In: CVPR. pp. 95–104. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.18 5
4. Busto, P.P., Gall, J.: Open Set Domain Adaptation. In: ICCV. pp. 754–763. IEEE, Venice (Oct 2017). https://doi.org/10.1109/ICCV.2017.88 4
5. Cao, Z., Long, M., Wang, J., Jordan, M.I.: Partial Transfer Learning with Selective Adversarial Networks. In: CVPR. pp. 2724–2732. IEEE, Salt Lake City, UT (Jun 2018). https://doi.org/10.1109/CVPR.2018.00288 4
6. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: CVPRW. pp. 702–703 (June 2020). https://doi.org/10.1109/cvprw50498.2020.00359 12
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848 11
8. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(59), 1–35 (2016) 4, 5, 10, 11
9. Gebru, T., Hoffman, J., Fei-Fei, L.: Fine-Grained Recognition in the Wild: A Multi-task Domain Adaptation Approach. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 1358–1367. IEEE, Venice (Oct 2017). https://doi.org/10.1109/ICCV.2017.151 2, 5
10. Ghifary, M., Kleijn, W.B., Zhang, M.: Domain Adaptive Neural Networks for Object Recognition. In: Pham, D.N., Park, S.B. (eds.) PRICAI 2014: Trends in Artificial Intelligence, vol. 8862, pp. 898–904. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-13560-1_76, series Title: Lecture Notes in Computer Science 5
11. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: ICML. pp. 513–520 (2011) 5
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020). https://doi.org/10.1145/3422622 7
13. Hedegaard, L., Sheikh-Omar, O.A., Iosifidis, A.: Supervised Domain Adaptation: A Graph Embedding Perspective and a Rectified Experimental Protocol. IEEE TIP **30**, 8619–8631 (2021). https://doi.org/10.1109/TIP.2021.3118978 4
14. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. In: ICML. pp. 1989–1998 (2018) 5, 10, 11
15. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: NeurIPS. pp. 18661–18673 (2020). https://doi.org/10.48550/arXiv.2004.11362 3, 4, 6, 9, 10

16. Koniusz, P., Tas, Y., Porikli, F.: Domain Adaptation by Mixture of Alignments of Second-or Higher-Order Scatter Tensors. In: CVPR. pp. 7139–7148. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.755 4
17. Li, Y., Wang, N., Shi, J., Hou, X., Liu, J.: Adaptive Batch Normalization for practical domain adaptation. PR **80**, 109–117 (Aug 2018). https://doi.org/10.1016/j.patcog.2018.03.005 5
18. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) NeurIPS. vol. 29. Curran Associates, Inc. (2016), https://proceedings.neurips.cc/paper_files/paper/2016/file/502e4a16930e414107ee22b6198c578f-Paper.pdf 4
19. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR. pp. 11976–11986 (2002). https://doi.org/10.1109/cvpr52688.2022.01167 11
20. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML. pp. 97–105 (2015) 5, 10, 11, 14
21. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts (2017). https://doi.org/10.48550/arXiv.1608.03983 12
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019). https://doi.org/10.48550/arXiv.1711.05101 12
23. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft (2013). https://doi.org/10.48550/arXiv.1306.5151 14
24. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV. pp. 2794–2802 (Oct 2017). https://doi.org/10.1109/iccv.2017.304 12
25. Motiian, S., Jones, Q., Iranmanesh, S., Doretto, G.: Few-shot adversarial domain adaptation. In: NeurIPS. vol. 30 (2017) 4, 6
26. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: ICCV. pp. 5715–5725 (2017). https://doi.org/10.1109/iccv.2017.609 4, 6
27. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019) 11
28. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open Set Domain Adaptation by Backpropagation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV, vol. 11209, pp. 156–171. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01228-1_10, series Title: Lecture Notes in Computer Science 4
29. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682 8
30. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: CVPR. pp. 2107–2116 (July 2017). https://doi.org/10.1109/cvpr.2017.241 12
31. Sielemann, A., Wolf, S., Roschani, M., Ziehn, J., Beyerer, J.: Synset Boulevard: A Synthetic Image Dataset for VMMR. In: 2024 IEEE International Conference on Robotics and Automation (ICRA) (2024) 3, 11
32. Sochor, J., Špaňhel, J., Herout, A.: Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance. IEEE Transactions on Intelligent Transportation Systems **PP**(99), 1–12 (2018). https://doi.org/10.1109/TITS.2018.2799228 2

33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (June 2016). https://doi.org/10.1109/cvpr.2016.308 12
34. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous Deep Transfer Across Domains and Tasks. In: ICCV. pp. 4068–4076. IEEE, Santiago, Chile (Dec 2015). https://doi.org/10.1109/ICCV.2015.463 4, 5
35. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial Discriminative Domain Adaptation. In: CVPR. pp. 2962–2971. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.316 4
36. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312**, 135–153 (Oct 2018). https://doi.org/10.1016/j.neucom.2018.05.083 4
37. Wang, S., Chen, X., Wang, Y., Long, M., Wang, J.: Progressive adversarial networks for fine-grained domain adaptation. In: CVPR. pp. 9213–9222 (June 2020). https://doi.org/10.1109/cvpr42600.2020.00923 2, 5, 14
38. Wang, Y., Song, R.J., Wei, X.S., Zhang, L.: An Adversarial Domain Adaptation Network For Cross-Domain Fine-Grained Recognition. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1217–1225. IEEE, Snowmass Village, CO, USA (Mar 2020). https://doi.org/10.1109/WACV45572.2020.9093306 2, 5
39. Yang, L., Luo, P., Change Loy, C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: CVPR. pp. 3973–3981 (June 2015). https://doi.org/10.1109/cvpr.2015.7299023 2, 11, 14
40. Yao, T., Yingwei Pan, Ngo, C.W., Houqiang Li, Tao Mei: Semi-supervised Domain Adaptation with Subspace Learning for visual recognition. In: CVPR. pp. 2142–2150. IEEE, Boston, MA, USA (Jun 2015). https://doi.org/10.1109/CVPR.2015.7298826 4
41. You, K., Long, M., Cao, Z., Wang, J., Jordan, M.I.: Universal Domain Adaptation. In: CVPR. pp. 2715–2724. IEEE, Long Beach, CA, USA (Jun 2019). https://doi.org/10.1109/CVPR.2019.00283 4, 6
42. Yu, H., Jiang, R., Li, A.: Striking a Balance in Unsupervised Fine-Grained Domain Adaptation Using Adversarial Learning. In: Li, G., Shen, H.T., Yuan, Y., Wang, X., Liu, H., Zhao, X. (eds.) Knowledge Science, Engineering and Management, vol. 12275, pp. 401–413. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-55393-7_36, series Title: Lecture Notes in Computer Science 2, 5
43. Yu, Q., Hashimoto, A., Ushiku, Y.: Divergence optimization for noisy universal domain adaptation. In: CVPR. pp. 2515–2524 (2021) 5
44. Zhang, J., Ding, Z., Li, W., Ogunbona, P.: Importance Weighted Adversarial Nets for Partial Domain Adaptation. In: CVPR. pp. 8156–8164. IEEE, Salt Lake City, UT (Jun 2018). https://doi.org/10.1109/CVPR.2018.00851 4
45. Zhang, X., Yu, F.X., Chang, S.F., Wang, S.: Deep Transfer Network: Unsupervised Domain Adaptation (2015). https://doi.org/10.48550/ARXIV.1503.00591 5
46. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017). https://doi.org/10.1109/iccv.2017.244 3, 5, 6, 7, 10, 11, 12