Few-Shot Supervised Prototype Alignment for Pedestrian Detection on Fisheye Images

Thaddäus Wiedemer[1,2,3], Stefan Wolf[2,3,4], Arne Schumann[3,4], Kaisheng Ma[1], Jürgen Beyerer[3,2]
[1]Tsinghua University, [2]Karlsruhe Institute of Technology,
[3]Fraunhofer IOSB, [4]Fraunhofer Center for Machine Learning
thaddaeus.wiedemer@uni-tuebingen.de, stefan.wolf@iosb.fraunhofer.de

Abstract

Complete and pre-trained models are readily available for download for object detection and can perform well on datasets containing everyday images. Domain adaptation is used to transfer models to more specific datasets with characteristics not present in pre-training. We propose the novel adaptation setting of pedestrian detection in fisheye images, where target samples are scarce but annotated. Our setting provides interesting new challenges for adaptation due to global perspective changes and geometric distortions not found in existing adaptation tasks. To this end, we introduce loss coupling for unsupervised adversarial adaptation and boost prototype-based adaptation with ground-truth information. We additionally propose a novel supervised adaptation head for features in the bounding box regressor. Our method leads to more stable adversarial training and outperforms supervised and unsupervised baselines. Our method requires half the amount of training samples for small datasets to achieve the same performance as supervised fine-tuning.

1. Introduction

Object detection is a well-studied field of research, and many recent machine learning models achieve excellent accuracy through the models' capacity to learn from large amounts of labeled data. However, in more specialized domains, the type of data encountered can differ significantly from existing large training datasets.

An essential and challenging domain of this nature is the footage of omnidirectional cameras, called fisheye images, which is common in surveillance and security applications [3–6]. Fisheye cameras are widely used since they offer an ultra-wide-angle field of view with a single sensor and are thus cost-effective in covering large areas. Large fisheye datasets, *e.g.* for person detection, are unavailable, as data protection laws and security concerns make it challenging to collect large amounts of data, especially for public datasets. Fortunately, large amounts of annotated pedestrian imagery recorded by conventional consumer cameras exist. Knowledge from these images can be transferred across the domain gap to the fisheye setting by suitable domain adaptation methods.

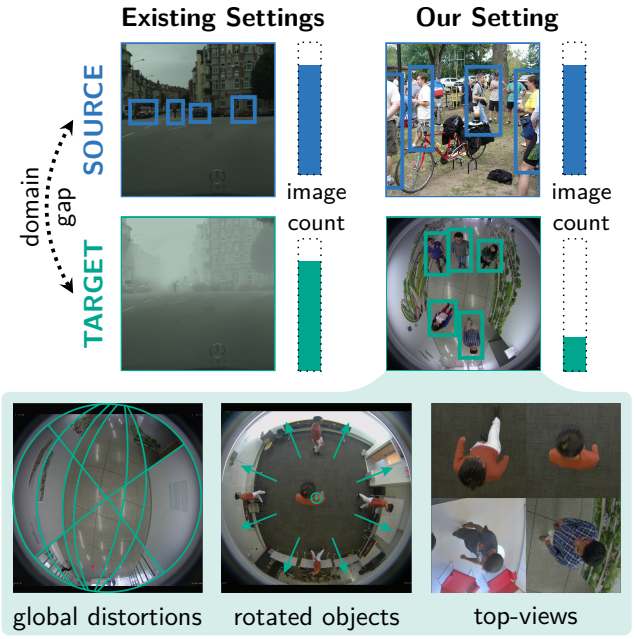

Figure 1. The proposed adaptation setting featuring a small, annotated target dataset. Existing domain-adaptive object detection settings instead use large unannotated datasets. Our setting extends the usual domain gap consisting of pixel-level color or brightness changes (*e.g.* through added fog [1]) or class-level appearance changes (*e.g.* in clip art images [2]) by global distortions, arbitrary object rotations, and a changed viewpoint.

Similar to other domain adaptation tasks, our target domain differs from the source domain due to scene differences and changes in the appearance of classes. However, note that the domain gap is further widened by the unique camera geometry, which results in the distortion and rota-

tion of objects, as shown in Figure 1. These properties also render the synthesis of artificial training samples by visual adaptation of source domain samples similar to Arruda *et al.* [7] or Lee *et al.* [8] infeasible, making the small number of training images an even bigger challenge.

However, the smaller amount of data makes it feasible to annotate all images resulting in a different domain adaptation setting rarely considered in prior works: few-shot supervised domain adaptation [9]. We propose a novel approach for this setting with fisheye images as the target domain. Specifically, our contributions can be summarized as follows.

- We introduce a novel and challenging domain adaptation task featuring global distortions and perspective changes. We publish this setting, including our dataset partitioning and code[1], to encourage further research on this domain.
- We propose a novel loss coupling scheme for adversarial adaptation that provides a more stable training process and improves adaptation results.
- To incorporate ground-truth information in the adaptation mechanism, we propose two novel supervised prototype-based adaptation heads for fine-grained adaptation of classification and bounding box regression.
- We thoroughly analyze the effect of dataset size on model performance resulting in new insights for researchers and practitioners. To the best of our knowledge, the effect of training set size in domain-adaptive object detection has not yet been analyzed to this extent.

Our experiments show that the proposed method requires 2.1 times fewer training samples to perform on par with supervised fine-tuning training.

2. Related work

Object detection and domain adaptation are well-established fields with a broad range of literature. This section summarizes existing research in both domains focusing on fisheye pedestrian detection and few-shot domain adaptive object detection.

Object detection. The two main branches of development for object detection in the past decade are single-stage architectures [10–13] and two-stage architectures [14–16]. The former branch predicts detections in a single pass, resulting in a simpler architecture. Recently, anchor-free approaches for one-stage architectures were proposed by Tian *et al.* [17] and Kong *et al.* [18] to overcome the need for choosing suitable anchor box sizes. Two-stage architectures uses a refinement stage to predict classes (if needed) and refine bounding boxes yielding more accurate predictions. Novel methods like DETR [19] refrain from using a pixel-based prediction approach. Instead, they use a transformer architecture to directly predict a list of objects rendering the need for post-processing steps superfluous.

Multiple adjustments have been proposed to improve detection performance on fisheye images. The majority of works use a pre-trained detector and apply fine-tuning and additional data augmentation to adjust to fisheye images [4, 20–23]. In this regard, object detectors have been extended to incorporate depth information [4, 22], location-specific refinement of bounding boxes [22, 23], and random rotation augmentation [23]. Another line of works focuses on transforming image patches [24–29] or extracted images features [6, 30] to reduce distortions. On the architectural side, convolutional neural networks (CNNs) have been adapted to work on spherical data by distributing the convolution sample locations on a sphere and using a rotation-invariant convolution operation [31–34]. An important design decision in many works is the modality of the output. While most object detection approaches use axis-aligned rectangular bounding boxes, the use of shapes similar to circular sectors [21], bounding ellipses [22], and oriented bounding boxes [23, 28, 29] have been investigated to fit the shape of people in fisheye images more accurately.

Domain adaptation. Following Wang *et al.* [35] and Li *et al.* [36], methods for domain adaptation can be sorted into three categories. All approaches aim to align intermediate features across both domains. (1) **Discrepancy-based** methods achieve alignment through optimizing an explicit loss function. Typical criteria for the optimization are classification [37, 38] (fine-tuning can be thought of as a special case of this), population statistics [39], architectural considerations [40], and geometric properties [41]. (2) **Adversarial-based** approaches align features by training an adversarial domain discriminator and encouraging domain-confusion [42]. (3) **Reconstruction-based** methods use autoenconders [43] or cyclic generative adversarial networks (GANs) [44] to generate synthetic training samples [7, 8, 45–47].

While most approaches were initially proposed for classification tasks, many have been adapted for domain-adaptive object detection. The adapted approaches include disrepancy-based methods [48–53] and adversarial-based methods [54–60] as well as reconstruction-based approaches [7, 8, 46, 47, 61, 62] using CycleGAN [44] or AugGAN [63]. Multiple authors investigated the combination of adversarial-based approaches with either discrepancy-based [2, 64–68] or reconstruction-based methods [69].

[1] https://github.com/ThaddaeusWiedemer/FisheyeSPA

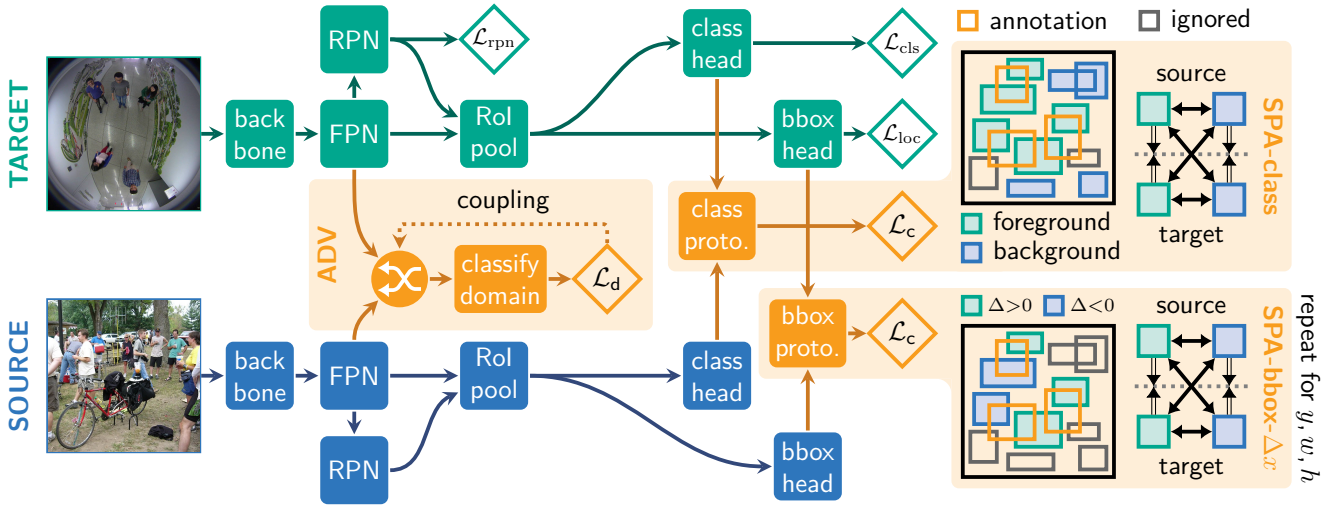

Figure 2. We apply domain adaptation to both network stages: Adversarial adaptation (ADV) on global image features and supervised prototype-based adaptation (SPA) on region-specific features. For ADV, we propose to couple the domain discriminator's influence to its loss, resulting in higher precision and more stable training. In SPA, we compute each region's overlap with annotations to build foreground/background prototypes to align classification features (SPA-class). In the bounding box regressor, we form prototypes from regions with positive/negative regression targets (SPA-bbox, depicted is the assignment of regions to prototypes based on the offset Δx of their center points). In both cases, regions with an overlap smaller than some threshold are ignored (grey). Prototypes of the same category are aligned across domains, while different categories are kept distinguishable. $\mathcal{L}_{\text{rpn}}$, $\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{loc}}$ are the training losses of Faster R-CNN.

Most previous works investigate an unsupervised domain adaptation scenario that lacks class and bounding box labels in the target domain. Few-shot supervised domain adaptation with a small number of annotated images in the target domain, as explored in this work, has only been rarely investigated [9]. This scenario clearly differs from traditional few-shot object detection which detects novel classes in the same domain [70–72].

Dissociation. To the best of our knowledge, domain adaptive object detection targeting fisheye images with its specific challenges introduced in Section 1 has not yet been investigated, neither in an unsupervised nor in a few-shot supervised scenario. Methodologically, the most related approaches are Wang *et al.* [9] and Xu *et al.* [51]. Compared to Xu *et al.* [51], we introduce the exploitation of ground-truth information in multiple ways and a novel adaptation head called SPA-bbox for bounding box regression. Moreover, we apply image-level adversarial-based adaptation. Compared to Wang *et al.* [9], we introduce a novel loss coupling for adversarial adaptation and apply our novel discrepancy-based method called supervised prototype alignment (SPA) for instance-level adaptation. Compared to all prior works on fisheye pedestrian detection, our method is completely domain agnostic. It does not rely on available camera parameters or assumptions on the exact nature of distortions.

3. Method

We propose a novel approach to include supervision in domain adaptation for few-shot settings. The complete method is depicted in Figure 2. Faster R-CNN [14] serves as the base detection architecture for all experiments since the two-stage architecture facilitates alignment of features in all parts of the detection pipeline. In contrast to unsupervised adaptation methods, we fine-tune the detector on the target domain but ignore training losses on the source domain. As in previous works [9], adaptation is performed on global, *image-level* features in the first stage of the network and local, region-specific, *instance-level* features in the second stage.

Image-level features are aligned using unsupervised adversarial domain adaptation [42] which we modify for improved stability on small datasets in Section 3.1. We propose supervised prototype alignment (SPA) for instance-level alignment of features corresponding to regions of interest (RoIs) in Section 3.2. Prototypes are generated and aligned by class (SPA-class) and bounding box offsets (SPA-bbox).

3.1. Image-level adversarial adaptation

We first align extracted global image features between domains through a domain discriminator on feature patches similar to Wang *et al.* [9] and Zhu *et al.* [44]. Aligning small patches reduces the number of parameters in the discriminator and guarantees a fixed-size input independent of feature

map size. In contrast to Wang *et al.* [9], we simplify patch extraction by choosing a fixed spatial size $s \times s$ and sample n random patch locations

$$(x_i, y_i) \sim (\mathcal{U}_w, \mathcal{U}_h) \tag{1}$$

for each image in the training batch, where $\mathcal{U}_k$ denotes a uniform distribution on $\left[\frac{s}{2}, k - \frac{s}{2}\right]$. Using fixed dimensions and a single square aspect ratio has the advantage that no subsequent pooling is required to process patches, preserving high-frequency information. Patches are extracted from the feature map in the feature pyramid network (FPN) [73] with the highest spatial resolution, referred to as `neck 0`. This feature map provides a good trade-off between rich semantic information and high spatial resolution and is highly interconnected to other FPN features. We experiment with aligning other FPN or backbone features—or even multiple global image features simultaneously—but find adaptation on only `neck 0` to yield the best results.

Loss coupling. Adversarial learning commonly suffers from unstable training, which is made worse by the scarcity of training samples. The coupling of the domain discriminator to the main network is commonly implemented according to Ganin *et al.* [42] as a gradient reverse layer $R(\boldsymbol{x})$ where the forward pass is the identity function, and the backward pass is defined as

$$\frac{\partial R}{\partial \boldsymbol{x}} = -\lambda \boldsymbol{I}. \tag{2}$$

The additional factor $\lambda \in [0, 1]$ can be thought of as the discriminator's influence on the main network. Ganin *et al.* [42] increase the influence monotonously over all N training iterations using

$$\lambda(i) = \frac{2}{1 + \exp\left(-\gamma \cdot \frac{i}{N}\right)} - 1 \tag{3}$$

with γ usually set to 10. As a result, the adversarial loss does not affect the main network until the discriminator is initialized. However, this schedule is ineffective in few-shot settings. It implies a monotonous increase in the quality of the discriminator throughout training, which cannot be guaranteed with limited update steps per epoch. Instead of using a fixed schedule, we propose to couple the influence in each iteration i to the discriminator's quality, using the domain classification loss as a proxy. Coupling the influence to the loss reduces the impact on the main network when discrimination performance is poor, leading to more stable training. The coupling is implemented as

$$\lambda(i) = \exp\big(-\mathcal{L}_\text{d}(i-1)\big) \tag{4}$$

with an exponential function mapping the discriminator's negative log-likelihood loss $\mathcal{L}_\text{d}(i) \in [0, \infty)$ to the influence factor $\lambda \in (0, 1]$. Our coupling scheme also simplifies adversarial adaptation by removing the hyperparameter γ in the original formulation.

3.2. Instance-level prototype-based adaptation

Features after RoI pooling in the second stage of Faster R-CNN are aligned using prototypes and a contrastive loss [74]. A prototype is the average feature representation of some category, obtained by clustering and aggregating features according to that category.

Unsupervised alignment. We first implement Graph-based Prototype Alignment (**GPA**) based on the code provided by Xu *et al.* [51]. Input features are embedded into a lower-dimensional space using an FC layer. Region coordinates are used to compute an adjacency matrix

$$A_{ij} = \text{IoU}(r_i, r_j) \tag{5}$$

of all proposed regions r_i, r_j. Features $\boldsymbol{F} = (\boldsymbol{f}_0, ..., \boldsymbol{f}_R)$ and class predictions $\boldsymbol{p}^{(k)} = \left(p_0^{(k)}, ..., p_R^{(k)}\right)$ for all R regions are then aggregated into *instance prototypes*

$$\tilde{\boldsymbol{F}} = \boldsymbol{A}\boldsymbol{F} \quad \text{and} \tag{6}$$

$$\tilde{\boldsymbol{p}}^{(k)} = \boldsymbol{A}\boldsymbol{p}^{(k)} \tag{7}$$

which can be thought of as the average feature of each object instance in the image. The aggregation is necessary since one object might be partly represented in several different regions. The *class prototypes*

$$\boldsymbol{c}^{(k)} = \frac{\sum_{i=1}^{N} \tilde{\boldsymbol{p}}_i^{(k)} \cdot \tilde{\boldsymbol{F}}_i^T}{\sum_{i=1}^{N} \tilde{\boldsymbol{p}}_i^{(k)}} \tag{8}$$

aggregate instance prototypes over all N images in the batch for all k classes. In this formulation, regions are implicitly assigned to prototypes based on the model's classification of each region. Therefore, the actual class labels are not needed.

Finally, a *contrastive loss* $\mathcal{L}_\text{c} = \mathcal{L}_\text{intra} + \mathcal{L}_\text{inter}$ with an intra-class component and an inter-class component as proposed by Hadsell *et al.* [74] is used to align the prototypes. The intra-class loss can in principle be implemented as any D-dimensional distance metric d. We follow Xu *et al.* [51] and use the mean-squared distance, resulting in

$$\mathcal{L}_\text{intra} = d\left(\tilde{\boldsymbol{F}}^{(\mathcal{S})}, \tilde{\boldsymbol{F}}^{(\mathcal{T})}\right) = \frac{1}{D} \sum_i \left|\tilde{F}_i^{(\mathcal{S})} - \tilde{F}_i^{(\mathcal{T})}\right|^2. \tag{9}$$

The inter-class loss

$$\mathcal{L}_\text{inter} = \left(\frac{m - \sqrt{d}}{m}\right)^2 \cdot \left(\max\left\{0, m - \sqrt{d}\right\}\right)^2 \tag{10}$$

encourages the model to increase the distance between classes up to a margin parameter $m = 1$. Minimizing the loss pair aligns prototypes of the same class across both domains, keeping classes across domains and within each domain distinguishable.

Supervised alignment by class. The implicit assignment of regions to classes can be made explicit when class labels are available. To this end, we propose Supervised Prototype Alignment (**SPA**). Regions r_i overlapping more than some threshold t_f with any ground truth are considered foreground samples; regions overlapping less than some threshold t_b with all ground truths are considered background samples. A region's maximum overlap

$$a_i = \max_j \{\text{IoU}(r_i, g_j)\} \tag{11}$$

with any ground-truth instance g_j is used to calculate the weights

$$a_i^{(f)} = \max\{a_i - t_f, 0\} + t_f \quad \text{and} \tag{12}$$

$$a_i^{(b)} = 1 - \min\{a_i - t_b, 0\} + t_b \tag{13}$$

of features belonging to both classes. The extension to more than two classes is trivial but not needed for the setting of this work. The class prototypes

$$\boldsymbol{c}^{(k)} = \frac{\sum_i^N \boldsymbol{a}_i^{(k)} \boldsymbol{F}_i}{\sum_i^N \boldsymbol{a}_i^{(k)}} \quad \text{with} \quad k \in \{f, b\} \tag{14}$$

aggregate all region proposals over N images and can be aligned with the contrastive intra-class and inter-class loss pair, just as in GPA.

Supervised alignment by bounding box offset. So far, prototypes are only generated and aligned for different classes. This intuitively improves the classification task of object detection but does not directly benefit the localization task. If features were perfectly aligned by class—and only by class—regression of different bounding box offsets Δx, Δy, Δw, Δh for objects of the same class would be impossible. Xu *et al.* [51] use prototype alignment in a shared head of the second stage, leaving only a single FC layer to regress bounding box dimensions from class-aligned features. On the shared features, the choice is between improving classification or localization accuracy. It makes sense to focus on classification since completely undetected objects (objects predicted to belong to the background) are worse than slightly mislocalized ones. But the contrastive loss pair from equations 9 and 10 has no limitation to align prototypes only by class. Therefore, we propose to align features separately in the classification head and bounding box regressor based on prototypes useful for each task.

Our proposed **SPA-bbox** follows the same idea as **SPA-class** introduced above. Regions are assigned to the ground truth with which they overlap the most. Only regions with an IoU greater than a threshold t are considered. Instead of assigning regions to foreground prototypes and background prototypes, their offsets along each bounding box offset dimension are computed. For each dimension d, features with a positive or negative offset Δd are aggregated into *offset prototypes*

$$\boldsymbol{p}^{(dk)} = \frac{\sum_i^N \boldsymbol{a}_i^{(dk)} \boldsymbol{F_i}}{\sum_i^N \boldsymbol{a}_i^{(dk)}},\ a_{i_j}^{(dk)} \in \{0, 1\},\ k \in \{p, n\}. \tag{15}$$

The resulting four pairs of positive/negative prototypes along different offset dimensions are not mutually exclusive (*e.g.* $\Delta x < 0$ and $\Delta y > 0$ can be valid for the same RoI). Therefore prototypes along different dimensions should not be separated using the contrastive inter-category loss. Instead, an individual contrastive loss is defined for each dimension. This is similar to using four separate SPA-bbox heads, each for another offset dimension. The only difference is that all prototypes are generated from the same feature embeddings (the same initial FC layer is used for all dimensions). Each loss pair's influence is reduced by a factor of 0.25.

4. Experiments

Datasets. We use **PIROPO** [75] and **Mirror Worlds** [76] as target datasets due to their large size compared to other fisheye datasets. Both datasets contain indoor recordings of people sitting, walking, standing, and interacting in multiple rooms. Labels in the form of *axis-aligned rectangular bounding boxes* are available for a small number of training frames (2357 and 819) and test frames (357 and 481) thanks to Tamura *et al.* [23]. Available *oriented rectangular bounding boxes* are not used in our setting as rotation information provides little benefit to surveillance settings and can be inferred from an object's location. To evaluate the relationship between number of training images and model performance, random subsets of sizes $\{1b, 2b, 5b\}$ with $b \in \{1, 10, 100, 1000\}$ are generated from both training sets (sampled evenly across cameras and rooms). Results for each size are averaged over the same three random instances `a`, `b`, `c` for each size. A custom subset of COCO 2017 [77] called **COCO-person** and containing 64k images where only pedestrians are annotated is used as the source domain.

Baselines. We compare our method against four baselines: (1) The model trained on **source data only** without additional fine-tuning. (2) **Fine-tuned model** without additional adaptation. (3) The model trained only with patch-based

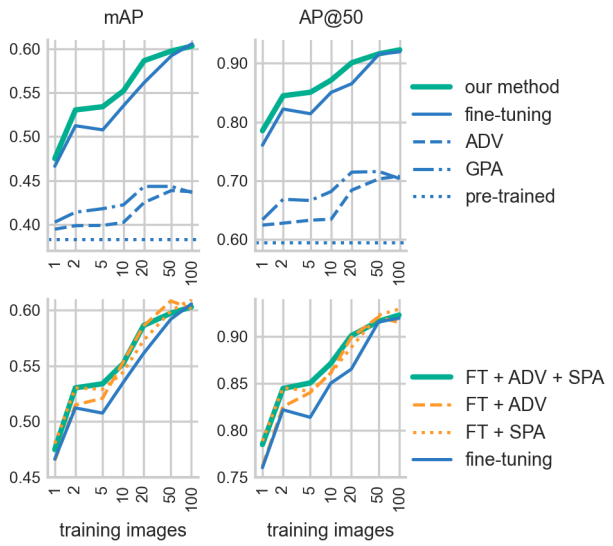

Figure 3. Results of our combined model compared against the fine-tuning baseline, unsupervised adversarial adaptation (ADV), and unsupervised prototype-based adaptation (GPA) (top row). Both unsupervised methods achieve a better accuracy than a model solely pre-trained on COCO, but are outperformed by fine-tuning with even a single image. Our supervised domain adaptation boosts fine-tuning performance across nearly all dataset sizes. We also compare our final model to its components (bottom row). Supervised prototype alignment (SPA; including SPA-class and SPA-bbox) improves performances on small dataset sizes, while adversarial adaptation works better on larger ones. The combined method performs best on datasets with less than 100 images.

unsupervised adversarial adaptation (ADV). This baseline still uses our proposed loss-coupling scheme. (4) The model trained only with **unsupervised prototype-based adaptation**. This corresponds closely to GPA [51] but uses our modified training setting, different loss weights, and a split R-CNN head. We cannot compare directly to the supervised few-shot approach by Wang *et al.* [9] due to the lack of code and the description being insufficient for reimplementation.

Implementation. The method is built on the open source frameworks **MMDetection** 2.12.0 [78] and **MMCV** 1.3.6 [79] which also provide a pre-trained version of **Faster R-CNN** [14] (36 epochs on **COCO** [77] and 12 epochs on **COCO-person**). We modify the model to use separate heads for classification and bounding box regression by duplicating the previously shared layers. Prototype alignment is based on the code provided by Xu *et al.* [51]. To stabilize training, we re-balance all intra-class and inter-class loss terms with a factor of 10 and 0.1, respectively. We empirically determined that setting the SPA-class thresholds t_f and t_b to 0.75 and 0.25 works well in practice, but the exact values do not have a large impact on performance.

	Number of training images		
Method	1	10	100
ADV	$0.624_{\pm 0.014}$	$0.634_{\pm 0.027}$	$0.707_{\pm 0.038}$
GPA	$0.633_{\pm 0.053}$	$0.681_{\pm 0.013}$	$0.703_{\pm 0.038}$
fine-tuning	$0.760_{\pm 0.038}$	$0.850_{\pm 0.018}$	$0.920_{\pm 0.008}$
FT + ADV	$0.762_{\pm 0.045}$	$0.861_{\pm 0.017}$	$0.915_{\pm 0.009}$
FT + SPA	$\mathbf{0.789}_{\pm 0.024}$	$0.861_{\pm 0.018}$	$\mathbf{0.929}_{\pm 0.004}$
combined	$0.785_{\pm 0.002}$	$\mathbf{0.871}_{\pm 0.031}$	$0.923_{\pm 0.001}$

Table 1. AP@50 on PIROPO of methods from Figure 3 for selected dataset sizes with 95 % confidence intervals. Exact values for all datapoints can be found in the supplementary materials.

Adversarial domain discriminators are implemented with 3 FC layers ($x \rightarrow 128 \rightarrow 32 \rightarrow 2$) and process $n = 32$ patches of size 35×35 per image. An evaluation of hyperparameters is included in the supplementary material. For more implementation details, please refer to our published code.

Training. As is usual for few-shot settings [9, 80, 81], the learning rate is fixed to 0.001 and warmup iterations are removed. We use SGD with a momentum of 0.9 and a weight decay of 0.0001. The training uses a total batch size of 16 on 4 NVIDIA GeForce GTX 1080Ti GPUs. For training sets with less than 16 or 4 images, the batch size and the number of GPUs are decreased accordingly. Images on both domains are resized to 800×800 pixels and normalized. Random rotation was tested and found to decrease performance, so only random horizontal flips with probability 0.5 are used for data augmentation. Models are trained with early stopping for 40 epochs, results are reported for the checkpoint achieving the highest AP@50 on the test set.

Metrics. We measure performance as COCO-style average precision (mAP, AP@75, AP@50) and log-average miss-rate (LAMR), which is more common in security-related settings. Note that mAP, AP@75, and AP@50 increase with improved detection performance, while LAMR decreases. We omit some of these metrics in diagrams and only show the mean over three model instances for visual clarity but report the full data with confidence intervals in the supplementary material. The three model instances are trained on different subsets of the dataset to compensate for randomness due to sampling. We quantify improvements over the baseline not only along the y-axis (as increase in percentage points) but also along the x-axis as the average distance between performance curves. We express this quantity as a factor that indicates how much larger the training set would have to be for the baseline to reach the same performance. Details on the derivation of this value can be found in the supplementary material.

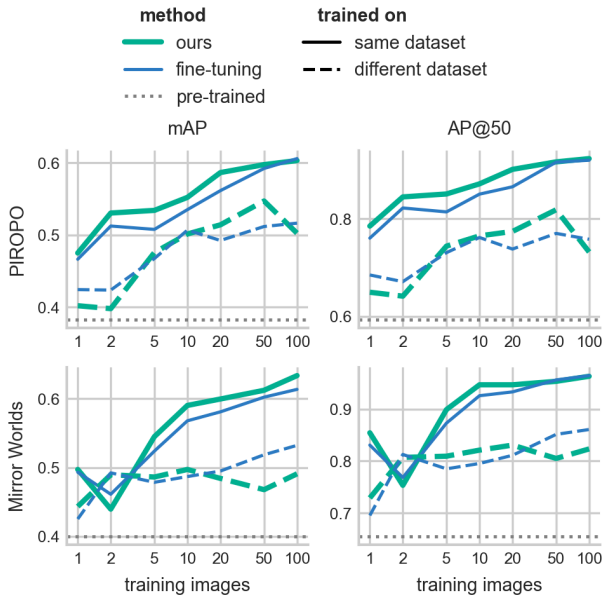

Figure 4. Performance of combined adversarial adaptation and SPA (blue) compared to a fine-tuned model on PIROPO (top row) and Mirror Worlds (bottom row). The performance of models trained and tested on the same dataset is shown as a solid line, while models transferred between datasets are indicated with dashed lines. The performance of the pre-trained model without additional training is shown dotted in gray.
The performance drop of transferred models clearly shows a remaining domain gap even between fisheye datasets. Adaptation in training can slightly increase performance across datasets, indicating that adaptation helps with regularization during training.

4.1. Quantitative results

Results on PIROPO. The top row of Figure 3 shows the performance of our final model, including all improvements evaluated on PIROPO [75] compared to a fine-tuning baseline, unsupervised domain adaptation, and a model solely pre-trained on COCO [77]. Our model outperforms the fine-tuning baseline across almost all evaluated dataset sizes. Only for the largest datasets with 100 images, the performances are tied. For AP@50, the performance boost averages 1.9 percentage points across all data set sizes. The baseline requires 2.1 as many training samples on average to yield the same performance. These results demonstrate that adaptation methods can compensate for the lack of data samples in small datasets, even when the training uses annotations.

While achieving better results than the model pre-trained solely on COCO [77], both baselines using unsupervised domain adaptation are less accurate than the fine-tuning baseline. The fact that fine-tuning on even a single annotated image results in higher precision than unsupervised domain adaptation highlights the importance of research regarding few-shot supervised domain adaptation scenarios. Our results also indicate that unsupervised domain adaptation shows little additional benefit for more than 50 images.

The bottom row of Figure 3 compares the combined method to both our enhanced adaptation methods separately. The combination of both adaptation methods achieves the best overall results on both metrics, while adversarial adaptation and SPA perform slightly better on individual dataset sizes. Exact values for all methods in Figure 3 for selected sizes are listed in Table 1 including 95 % confidence intervals.

Cross-dataset results. We also evaluate our model on Mirror Worlds [76] (see Figure 4). Additionally, we test performance across datasets by training on one dataset and testing on the other. For training and testing on Mirror Worlds, our method outperforms the fine-tuning baseline. Results are only worse for a dataset size of 2 images. However, the baseline also exhibits a performance drop here, indicating that the randomly sampled training samples for this size are likely of low quality. For Mirror Worlds→PIROPO, our method outperforms the baseline on 5 to 50 training samples. For PIROPO→Mirror Worlds, the range is 1 to 20 images. All methods outperform the model solely pre-trained on COCO [77]. The results indicate that the domain gap consists of two parts. The first part contains the general characteristics of fisheye images. The model initially learns to reduce this gap, leading to increased performance on a fisheye dataset it has not been trained on. The second part contains the properties specific to a particular dataset, like the layout of rooms and the position of each camera. With larger datasets and more training iterations, the model overfits on the training domain at the cost of poorer generalization to other domains. Since our model adapts more easily to a particular domain, it overfits earlier than the fine-tuning baseline. The model trained on PIROPO overfits before the one trained on Mirror Worlds since PIROPO contains fewer unique cameras (4 vs. 7) and fewer unique people (2 vs. 5).

4.2. Ablation studies

We validate the effectiveness of our method through ablation studies described in this section. Due to high computational resource requirements for experiments across all dataset sizes with multiple subsets per size, our ablation studies are performed on a single subset with 20 images.

Supervision of Prototype Alignment. In Table 2, the results of different supervision types in the instance-level alignment with GPA as described in Section 3.2 are shown. While no supervision shows the overall worst precision, supervising the alignment of classes proves to have the largest impact on AP@50. Supervising the bounding box regression alignment benefits mostly mAP and AP@75, which target a high localization accuracy. As expected, the super-

Align	GT	mAP↑	AP@75↑	AP@50↑	LAMR↓
class	✗	0.574	0.643	0.886	0.153
class	✓	0.579	0.642	**0.901**	0.138
Δx	✓	0.589	0.662	0.900	**0.137**
$\forall \Delta d$	✓	**0.591**	**0.666**	0.899	0.139

Table 2. Performance on PIROPO-20a for different adaptation methods in the bounding box head. The combined alignment of all bounding box dimensions yields the best overall performance, especially on mAP and AP@75 which put greater emphasis on correct localization. GT denotes whether ground-truth information was used.

λ	mAP↑	AP@75↑	AP@50↑	LAMR↓
Constant	0.521	0.573	0.842	0.217
Increasing	0.536	0.590	0.861	0.188
Coupled (ours)	**0.563**	**0.627**	**0.895**	**0.148**

Table 3. Performance for different schedules of the influence factor λ in adversarial adaptation on PIROPO-20a. Coupling the influence to the discriminator's loss boosts overall performance.

vision of the bounding box regressor is most helpful when applied to each regression dimension.

Loss coupling. The impact of loss coupling for adversarial training on the precision of the model is shown in Table 3. We compare performance to training with a monotonously increasing influence factor λ as proposed by Ganin *et al*. [42] and a fixed influence factor $\lambda = 1$. Figure 5 shows the AP@50 over the course of 40 training epochs for all three λ-schedules. Training with a constant or increasing influence factor suffers from drastic performance drops, which require several epochs to recuperate from. This leads to decreased overall performance and is a problem in few-shot settings, where the small number of samples should be fully utilized and cannot be held back for validation and early-stopping. Our method leads to a more stable learning process with decreased performance drops and boosts overall performance. This is particularly valuable in real-world training settings without early-stopping, where stable training guarantees good performance independent of the exact number of training epochs.

5. Conclusion

This work presents the novel domain adaptation setting of pedestrian detection in fisheye images. The setting poses significantly different challenges than existing adaptation tasks due to geometric distortions, perspective changes, and the scarcity of training samples. We demonstrate that the combination of supervised domain adaptation with fine-tuning is effective in meeting these challenges. Our main contributions are stabilizing adversarial adaptation through

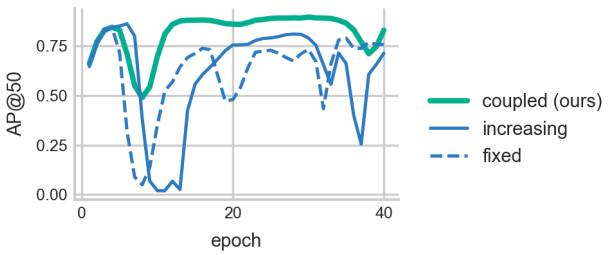

Figure 5. AP@50 for different schedules of the influence factor λ in adversarial adaptation on PIROPO-20a. Coupling the discriminator's influence to its loss leads to more stable training and better performance.

loss coupling and incorporating ground-truth information in prototype-based alignment. We align features for classification and localization by their respective regression targets, which leads to more fine-grained adaptation. On average, our method reduces the number of training samples needed to reach competitive performance by a factor of two compared to using fine-tuning. In real-world applications, this paves the way for quickly adapting cameras to their specific surroundings after installation.

Potential societal impact. Fisheye images are predominantly used in surveillance settings. While all methods in this area can be misused, our method does not extend to biometric data and does not facilitate the detection of particular groups of people. We think that the benefit to legitimate security applications outweighs the potential negative impact.

Limitations and future work. While experiments were only conducted on our newly proposed setting, all proposed methods are domain agnostic. Results should therefore generalize well to other supervised few-shot settings, including multi-class detection tasks. Similarly, our methods were tested only with a Faster R-CNN architecture. However, since they do not rely on mechanisms specific to this detector, we expect them to transfer well to other two-stage architectures. Our novel loss coupling scheme might be helpful on larger datasets as well. The proposed prototype-based alignment for bounding box features could be modified to use the model's predictions instead of ground-truth information. With this modification, it could be used for unsupervised adaptation. Future works might validate these expectations. Also left for future works is a deeper investigation of the data-versus-supervision trade-off. In our work, supervised adaptation is beneficial in addition to fine-tuning only on small datasets < 100 images. In this context, we also want to encourage the community to generally be more thorough in investigating the influence of dataset size on domain adaptation to give practitioners clear guidelines of how many samples are needed.

References

[1] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic Foggy Scene Understanding with Synthetic Data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, Sep. 2018. 1

[2] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 5001–5009. 1, 2

[3] H. Kim, E. Chae, G. Jo, and J. Paik, "Fisheye lens-based surveillance camera for wide field-of-view monitoring," in *2015 IEEE International Conference on Consumer Electronics (ICCE)*, 2015, pp. 505–506. 1

[4] N. Van Tuan, T. B. Nguyen, and S.-T. Chung, "ConvNets and AGMM based real-time human detection under fisheye camera for embedded surveillance," in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct. 2016, pp. 840–845. 1, 2

[5] H. Kim, J. Jung, and J. Paik, "Fisheye lens camera based surveillance system for wide field of view monitoring," *Optik*, vol. 127, no. 14, pp. 5636–5646, 2016. 1

[6] O. Krams and N. Kiryati, "People detection in top-view fisheye imaging," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2017, pp. 1–6. 1, 2

[7] V. F. Arruda, T. M. Paixão, R. F. Berriel, A. F. De Souza, C. Badue, N. Sebe, and T. Oliveira-Santos, "Cross-Domain Car Detection Using Unsupervised Image-to-Image Translation: From Day to Night," in *2019 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2019, pp. 1–8. 2

[8] H. Lee, M. Ra, and W.-Y. Kim, "Nighttime Data Augmentation Using GAN for Improving Blind-Spot Detection," *IEEE Access*, vol. 8, pp. 48 049–48 059, 2020. 2

[9] T. Wang, X. Zhang, L. Yuan, and J. Feng, "Few-Shot Adaptive Faster R-CNN," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 7166–7175. 2–4, 6

[10] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517–6525. 2

[11] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv: 1804.02767 [cs]*, Apr. 2018. 2

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, 2016, pp. 21–37. 2

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *arXiv:1708.02002 [cs]*, Feb. 2018. 2

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015. 2, 3, 6

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020. 2

[16] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving Into High Quality Object Detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 6154–6162. 2

[17] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9626–9635. 2

[18] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "Foveabox: Beyound anchor-based object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 7389–7398, 2020. 2

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020*, 2020, pp. 213–229. 2

[20] M. Saito, K. Kitaguchi, G. Kimura, and M. Hashimoto, "Human detection from fish-eye image by Bayesian combination of probabilistic appearance models," in *2010 IEEE International Conference on Systems, Man and Cybernetics*, Oct. 2010, pp. 243–248. 2

[21] I. Cinaroglu and Y. Bastanlar, "A Direct approach for human detection with catadioptric omnidirectional cameras," in *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, Apr. 2014, pp. 2275–2279. 2

[22] T. Wang, C.-W. Chang, and Y.-S. Wu, "Template-based people detection using a single downward-viewing fisheye camera," in *2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Nov. 2017, pp. 719–723. 2

[23] M. Tamura, S. Horiguchi, and T. Murakami, "Omnidirectional Pedestrian Detection by Rotation Invariant Training," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Annotations are provided for research purposes only, Jan. 2019, pp. 1989–1998. 2, 5

[24] A.-T. Chiang and Y. Wang, "Human detection in fish-eye images using HOG-based detectors over rotated windows," in *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, Jul. 2014, pp. 1–6. 2

[25] L. Meinel, M. Findeisen, M. Heß, A. Apitzsch, and G. Hirtz, "Automated real-time surveillance for ambient assisted living using an omnidirectional camera," in *2014 IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2014, pp. 396–399. 2

[26] M. Demirkus, L. Wang, M. Eschey, H. Kaestle, and F. Galasso, "People Detection in Fish-eye Top-views:" in *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2017, pp. 141–148. 2

[27] R. Seidel, A. Apitzsch, and G. Hirtz, "Improved Person Detection on Omnidirectional Images with Non-maxima Suppression," *arXiv:1805.08503 [cs]*, Mar. 2019. 2

[28] S.-H. Chiang, T. Wang, and Y.-F. Chen, "Efficient pedestrian detection in top-view fisheye images using compositions of perspective view patches," *Image and Vision Computing*, vol. 105, p. 104 069, Jan. 2021. 2

[29] T. Wang, Y.-Y. Hsieh, F.-W. Wong, and Y.-F. Chen, "Mask-RCNN Based People Detection Using A Top-View Fisheye Camera," in *2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, Nov. 2019, pp. 1–4. 2

[30] V. Srisamosorn, N. Kuwahara, A. Yamashita, T. Ogata, S. Shirafuji, and J. Ota, "Human position and head direction tracking in fisheye camera using randomized ferns and fisheye histograms of oriented gradients," *The Visual Computer*, vol. 36, no. 7, pp. 1443–1456, Jul. 2020. 2

[31] T. S. Cohen, M. Geiger, J. Koehler, and M. Welling, "Spherical CNNs," *arXiv:1801.10130 [cs, stat]*, Feb. 2018. 2

[32] B. Coors, A. P. Condurache, and A. Geiger, "SphereNet: Learning Spherical Representations for Detection and Classification in Omnidirectional Images," in *Computer Vision – ECCV 2018*, vol. 11213, 2018, pp. 525–541. 2

[33] M. Eder and J.-M. Frahm, "Convolutions on spherical images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2019. 2

[34] V. R. Kumar, S. A. Hiremath, M. Bach, S. Milz, C. Witt, C. Pinard, S. Yogamani, and P. Mäder, "FisheyeDistanceNet: Self-Supervised Scale-Aware Distance Estimation using Monocular Fisheye Camera for Autonomous Driving," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, May 2020, pp. 574–581. 2

[35] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018. 2

[36] W. Li, F. Li, Y. Luo, P. Wang, and J. sun, "Deep Domain Adaptive Object Detection: A Survey," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec. 2020, pp. 1808–1813. 2

[37] X. Zhang, F. X. Yu, S.-F. Chang, and S. Wang, "Deep transfer network: Unsupervised domain adaptation," *arXiv preprint arXiv:1503.00591*, 2015. 2

[38] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4068–4076. 2

[39] M. Ghifary, W. B. Kleijn, and M. Zhang, "Domain adaptive neural networks for object recognition," in *PRICAI 2014: Trends in Artificial Intelligence*, 2014, pp. 898–904. 2

[40] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, "Adaptive batch normalization for practical domain adaptation," *Pattern Recognition*, vol. 80, pp. 109–117, 2018. 2

[41] S. Chopra, S. Balakrishnan, and R. Gopalan, "Dlid: Deep learning for domain adaptation by interpolating between domains," in *ICML workshop on challenges in representation learning*, Citeseer, vol. 2, 2013. 2

[42] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning*, Jun. 2015, pp. 1180–1189. 2–4, 8

[43] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2551–2559. 2

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2242–2251. 2, 3

[45] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems*, vol. 29, 2016. 2

[46] C. Devaguptapu, N. Akolekar, M. M. Sharma, and V. N. Balasubramanian, "Borrow From Anywhere: Pseudo Multi-Modal Object Detection in Thermal Imagery," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2019, pp. 1029–1038. 2

[47] S. Liu, V. John, E. Blasch, Z. Liu, and Y. Huang, "IR2VI: Enhanced Night Environmental Perception by Unsupervised Thermal Image Translation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2018, pp. 1234–12 347. 2

[48] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. Macready, "A Robust Learning Approach to Domain Adaptive Object Detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 480–490. 2

[49] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring Object Relation in Mean Teacher for Cross-Domain Detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 11 449–11 458. 2

[50] Y. Cao, D. Guan, W. Huang, J. Yang, Y. Cao, and Y. Qiao, "Pedestrian detection with unsupervised multispectral feature learning using deep neural networks," *Information Fusion*, vol. 46, pp. 206–217, Mar. 2019. 2

[51] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-Domain Detection via Graph-Induced Prototype Alignment," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 12 352–12 361. 2–6

[52] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, "Exploring Categorical Regularization for Domain Adaptive Object Detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 11 721–11 730. 2

[53] A. Raj, V. P. Namboodiri, and T. Tuytelaars, "Subspace alignment based domain adaptation for rcnn detector," in *Proceedings of the British Machine Vision Conference (BMVC)*, Sep. 2015, pp. 166.1–166.11. 2

[54] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain Adaptive Faster R-CNN for Object Detection in the Wild," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 3339–3348. 2

[55] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-Weak Distribution Alignment for Adaptive Object Detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 6949–6958. 2

[56] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing Transferability and Discriminability for Adapting Object Detectors," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 8866–8875. 2

[57] Z. He and L. Zhang, "Multi-Adversarial Faster-RCNN for Unrestricted Object Detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 6667–6676. 2

[58] Z. Shen, H. Maheshwari, W. Yao, and M. Savvides, "SCL: Towards Accurate Domain Adaptive Object Detection via Gradient Detach Based Stacked Complementary Losses," *arXiv:1911.02559 [cs]*, Nov. 2019. 2

[59] C. Zhuang, X. Han, W. Huang, and M. Scott, "iFAN: Image-Instance Full Alignment Networks for Adaptive Object Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 13 122–13 129, Apr. 2020. 2

[60] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting Object Detectors via Selective Cross-Domain Alignment," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 687–696. 2

[61] C.-T. Lin, "Cross Domain Adaptation for on-Road Object Detection Using Multimodal Structure-Consistent Image-to-Image Translation," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 3029–3030. 2

[62] T. Guo, C. P. Huynh, and M. Solh, "Domain-Adaptive Pedestrian Detection in Thermal Images," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 1660–1664. 2

[63] S.-W. Huang, C.-T. Lin, S.-P. Chen, Y.-Y. Wu, P.-H. Hsu, and S.-H. Lai, "Auggan: Cross domain adaptation with gan-based data augmentation," in *Computer Vision – ECCV 2018*, 2018, pp. 731–744. 2

[64] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, "Diversify and Match: A Domain Adaptive Representation Learning Paradigm for Object Detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 12 448–12 457. 2

[65] S. Kim, J. Choi, T. Kim, and C. Kim, "Self-Training and Adversarial Background Regularization for Unsupervised Domain Adaptive One-Stage Object Detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 6091–6100. 2

[66] A. L. Rodriguez and K. Mikolajczyk, "Domain Adaptation for Object Detection via Style Consistency," *arXiv:1911.10033 [cs]*, Nov. 2019. 2

[67] Y. Zheng, D. Huang, S. Liu, and Y. Wang, "Cross-domain Object Detection through Coarse-to-Fine Feature Adaptation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 13 763–13 772. 2

[68] Y. Shan, W. F. Lu, and C. M. Chew, "Pixel and feature level based domain adaptation for object detection in autonomous driving," *Neurocomputing*, vol. 367, pp. 31–38, Nov. 2019. 2

[69] H.-K. Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H.-Y. Tseng, M. Singh, and M.-H. Yang, "Progressive domain adaptation for object detection," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 738–746. 2

[70] X. Wang, T. Huang, J. Gonzalez, T. Darrell, and F. Yu, "Frustratingly Simple Few-Shot Object Detection," in *International Conference on Machine Learning*, Nov. 2020, pp. 9919–9928. 3

[71] X. Dong, L. Zheng, F. Ma, Y. Yang, and D. Meng, "Few-Example Object Detection with Model Communication," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1641–1654, Jul. 2019. 3

[72] H. Chen, Y. Wang, G. Wang, and Y. Qiao, "LSTD: A Low-Shot Transfer Detector for Object Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. 3

[73] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," 2017, pp. 2117–2125. 4

[74] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality Reduction by Learning an Invariant Mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, Jun. 2006, pp. 1735–1742. 4

[75] C. R. del-Blanco, P. Carballeira, F. Jaureguizar, and N. García, "Robust people indoor localization with omnidirectional cameras using a Grid of Spatial-Aware Classifiers," *Signal Processing: Image Communication*, vol. 93, p. 116 135, Apr. 2021, The PIROPO database is available and free for research purposes. 5, 7

[76] *Mirror Worlds Challenge*, Dataset is licensed under Creative Commons BY-NC-SA. [Online]. Available: http://www2.icat.vt.edu/mirrorworlds/challenge/index.html. 5, 7

[77] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, Annotations are licensed under Creative Commons BY, Images must be used according to the Flickr Terms of Use, 2014, pp. 740–755. 5–7

[78] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open MMLab Detection Toolbox and Benchmark," *arXiv:1906.07155 [cs, eess]*, Jun. 2019, Source code is licensed under the Apache License 2.0. 6

[79] M. Contributors, *MMCV: OpenMMLab computer vision foundation*, Source code is licensed under the Apache License 2.0, 2018. [Online]. Available: https://github.com/open-mmlab/mmcv. 6

[80] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-Shot Object Detection With Attention-RPN and Multi-Relation Detector," 2020, pp. 4013–4022. 6

[81] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto, "Few-Shot Adversarial Domain Adaptation," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. 6