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

Thaddäus Wiedemer\textsuperscript{1,2,3}, Stefan Wolf\textsuperscript{2,3,4}, Arne Schumann\textsuperscript{3,4}, Kaisheng Ma\textsuperscript{1}, Jürgen Beyerer\textsuperscript{3,2}
\textsuperscript{1}Tsinghua University, \textsuperscript{2}Karlsruhe Institute of Technology, \textsuperscript{3}Fraunhofer IOSB, \textsuperscript{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.

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 autoencoders [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 discrepancy-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 $\Delta x$ 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. $L_{\text{rpn}}$, $L_{\text{cls}}$, $L_{\text{loc}}$ 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

\[
(x_i, y_i) \sim (U_w, U_h)
\]

for each image in the training batch, where \( U_w \) denotes a uniform distribution on \( \left[ \frac{1}{2}, k - \frac{1}{2} \right] \). 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 \( R(x) \) where the forward pass is the identity function, and the backward pass is defined as

\[
\frac{\partial R}{\partial x} = -\lambda I.
\]

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

\[
\lambda(i) = \frac{2}{1 + \exp(-\gamma \cdot \frac{i}{N})} - 1
\]

with \( \gamma \) 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

\[
\lambda(i) = \exp(-\mathcal{L}_d(i - 1))
\]

with an exponential function mapping the discriminator’s negative log-likelihood loss \( \mathcal{L}_d(i) \in [0, \infty) \) to the influence factor \( \lambda \in (0, 1] \). Our coupling scheme also simplifies adversarial adaptation by removing the hyperparameter \( \gamma \) 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

\[
A_{ij} = \text{IoU}(r_i, r_j)
\]

of all proposed regions \( r_i, r_j \). Features \( F = (f_0, \ldots, f_N) \) and class predictions \( p^{(k)} = (p_0^{(k)}, \ldots, p_R^{(k)}) \) for all \( R \) regions are then aggregated into instance prototypes

\[
\tilde{F}^i = A F \quad \text{and} \quad \tilde{p}^{(k)} = A p^{(k)}
\]

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

\[
\mathcal{C}^{(k)} = \frac{\sum_{i=1}^{N} \tilde{p}_i^{(k)} \cdot \tilde{F}_i^T}{\sum_{i=1}^{N} \tilde{p}_i^{(k)}}
\]

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* \( \mathcal{L}_c = \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}} \) 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

\[
\mathcal{L}_{\text{intra}} = d(\tilde{F}_i^{(S)}, \tilde{F}_j^{(T)}) = \frac{1}{D} \sum_i |\tilde{F}_i^{(S)} - \tilde{F}_i^{(T)}|^2.
\]

The inter-class loss

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

\[
a_i = \max_j \{ \text{IoU}(r_i, g_j) \}
\]

(11) with any ground-truth instance \( g_j \) is used to calculate the weights

\[
a_i^{(f)} = \max\{a_i - t_f, 0\} + t_f \quad \text{and} \quad a_i^{(b)} = 1 - \min\{a_i - t_b, 0\} + t_b
\]

(12) (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

\[
c^{(k)} = \frac{\sum_i^N a_i^{(k)} F_i}{\sum_i^N a_i^{(k)}} \quad \text{with} \quad k \in \{f, b\}
\]

(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 \( \Delta x \), \( \Delta y \), \( \Delta w \), \( \Delta h \) 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 \( \Delta d \) are aggregated into offset prototypes

\[
\rho^{(d_k)} = \frac{\sum_i^N a_i^{(d_k)} F_i}{\sum_i^N a_i^{(d_k)}} \quad a_i^{(d_k)} \in \{0, 1\}, \quad k \in \{p, n\}
\]

(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
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.

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>1</th>
<th>10</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADV</td>
<td>0.624 ± 0.014</td>
<td>0.634 ± 0.027</td>
<td>0.707 ± 0.038</td>
</tr>
<tr>
<td>GPA</td>
<td>0.633 ± 0.053</td>
<td>0.681 ± 0.013</td>
<td>0.703 ± 0.038</td>
</tr>
<tr>
<td>fine-tuning</td>
<td>0.760 ± 0.038</td>
<td>0.850 ± 0.018</td>
<td>0.920 ± 0.008</td>
</tr>
<tr>
<td>FT + ADV</td>
<td>0.762 ± 0.045</td>
<td>0.861 ± 0.017</td>
<td>0.915 ± 0.009</td>
</tr>
<tr>
<td>FT + SPA</td>
<td>0.789 ± 0.024</td>
<td>0.861 ± 0.018</td>
<td>0.929 ± 0.004</td>
</tr>
<tr>
<td>combined</td>
<td>0.785 ± 0.002</td>
<td>0.871 ± 0.031</td>
<td>0.923 ± 0.001</td>
</tr>
</tbody>
</table>

Adversarial domain discriminators are implemented with 3 FC layers ($x \rightarrow 128 \rightarrow 32 \rightarrow 2$) and process $n = 32$ patches of size $35 \times 35$ 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 $800 \times 800$ 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.
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-
Contributions are stabilizing adversarial adaptation through tuning, which is effective in meeting these challenges. Our main combination of supervised domain adaptation with fine-tuning guarantees good performance independent of the exact training settings without early-stopping, where stable training results in decreased performance drops and boosts overall performance. This is particularly valuable in real-world 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 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.

<table>
<thead>
<tr>
<th>Align</th>
<th>GT</th>
<th>mAP↑</th>
<th>AP@75↑</th>
<th>AP@50↑</th>
<th>LAMR↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>class</td>
<td>✗</td>
<td>0.574</td>
<td>0.643</td>
<td>0.886</td>
<td>0.153</td>
</tr>
<tr>
<td>class</td>
<td>✓</td>
<td>0.579</td>
<td>0.642</td>
<td>0.901</td>
<td>0.138</td>
</tr>
<tr>
<td>Δx</td>
<td>✓</td>
<td>0.589</td>
<td>0.662</td>
<td>0.900</td>
<td>0.137</td>
</tr>
<tr>
<td>∀Δd</td>
<td>✓</td>
<td>0.591</td>
<td>0.666</td>
<td>0.899</td>
<td>0.139</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>λ</th>
<th>mAP↑</th>
<th>AP@75↑</th>
<th>AP@50↑</th>
<th>LAMR↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.521</td>
<td>0.573</td>
<td>0.842</td>
<td>0.217</td>
</tr>
<tr>
<td>Increasing</td>
<td>0.536</td>
<td>0.590</td>
<td>0.861</td>
<td>0.188</td>
</tr>
<tr>
<td>Coupled (ours)</td>
<td><strong>0.563</strong></td>
<td><strong>0.627</strong></td>
<td><strong>0.895</strong></td>
<td><strong>0.148</strong></td>
</tr>
</tbody>
</table>

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.

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.
References


