Domain Adaptive Pedestrian Detection Based on Semantic Concepts

Patrick Feifel[1,2], Frank Bonarens[1] and Frank Köster[2,3]
[1]*Stellantis, Opel Automobile GmbH, Germany*
[2]*Carl von Ossietzky Universität Oldenburg, Germany*
[3]*Deutsches Zentrum für Luft- und Raumfahrt, Germany*

Keywords: Pedestrian Detection, Unsupervised Domain Adaptation, Interpretability.

Abstract: Pedestrian detection is subject to high complexity with a wide variety of pedestrian appearances and postures as well as environmental conditions. Building a sufficient real-world dataset is labor-intensive and costly. Thus, the application of synthetic data is promising, but deep neural networks show a lack of generalization when trained solely on synthetic data. In our work, we propose a novel method for concept-based domain adaptation for pedestrian detection (ConDA). In addition to the 2D bounding box prediction, an auxiliary body part segmentation exploits discriminative features of semantic concepts of pedestrians. Inspired by approaches to the inherent interpretability of DNNs, ConDA has been shown to strengthen generalization. This is done by enforcing a high intra-class concentration and inter-class separation of extracted body part features in the latent space. We report performance results regarding various training strategies, feature extractions and backbones for ConDA on the real-world CityPersons dataset.

1 INTRODUCTION

The reliable perception of vulnerable road users is a key requisite for automated driving. State-of-the-art deep neural networks for pedestrian detection (PD-DNNs) are specifically designed and developed for real datasets. A limiting factor is the labor-intensive and expensive manual generation of ground truth annotations for real-world images. Contrarily, synthetic data generation is more cost-effective, customizable and scalable. The consortium project KI Absicherung[1] (KI-A) focuses on generating synthetic data for pedestrian detection. Unfortunately, even photo-realistic synthetic data and real data exhibit a great domain shift. PD-DNNs solely trained with synthetic data generalize poorly to the real world. *Unsupervised domain adaptation* (UDA) aims to overcome insufficient generalization by enhancing supervised learning from synthetic data with unsupervised learning from real and unlabeled data. Due to the accessibility of synthetic data, additional task-specific labels such as body part segmentation (BPS) or instance segmentation can easily be provided. Based on this, pedestrian detection can be framed as a set of main tasks such as localization and classification and auxiliary segmentation.

Increasing emphasis on the segmentation of semantic concepts is also consistent with recent research on interpretability for DNNs. Interpretability is increasingly seen as a requirement for models used in safety-critical applications (Rudin, 2019). Some works analyze if an already trained DNN embeds predefined semantic concepts (Kim et al., 2018; Haselhoff et al., 2021). Another line of work focuses on the adaption of DNN architectures to enable *inherent interpretability* (Chen et al., 2019; Koh et al., 2020; Feifel et al., 2021b; Feifel et al., 2021a). The common idea is that classes are predicted based on distances to prototypes or concept vectors in the *latent space*. Methods for inherent interpretability make heavy use of achievements in the field of metric learning. The common goal is to diminish the disadvantages of the softmax loss and improve generalization.

Motivated by the interdependent benefits of inherent interpretability and metric learning, we propose a methodology for concept-based domain adaptation for pedestrian detection (ConDA). Our contribution can be summarized as follows:

- We show that learning an auxiliary body part segmentation from synthetic data improves generalization on real data of a DNN for pedestrian detection.
- We show that inherent interpretable DNNs inspired by techniques from metric learning offer superior generalization on real data.

[1]translation: AI Safeguarding, https://www.ki-absicherung-projekt.de/en/

Feifel, P., Bonarens, F. and Köster, F.
Domain Adaptive Pedestrian Detection Based on Semantic Concepts.
DOI: 10.5220/0011690900003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 652-659
ISBN: 978-989-758-634-7; ISSN: 2184-4321

- We propose a novel methodology for concept-based domain adaptive pedestrian detection (ConDA) that enhances supervised learning from synthetic data with unsupervised learning from real and unlabeled data.

2 RELATED WORK

2.1 Pedestrian Detection

Pedestrian detection is about locating and classifying 2D bounding boxes for pedestrians in a given image. Recent research has produced more and more high-performing anchor-free approaches. Two-stage PD-DNNs such as F2DNet (Khan et al., 2022) apply a region proposal as an intermediate step to identify possible areas of an image that might hold an object. Contrarily, CSP (Liu et al., 2019), APD (Zhang et al., 2020) and BGCNet (Li et al., 2020) use a simple one-stage architecture to achieve state-of-the-art performance. These PD-DNNs predict keypoints, most commonly the center point of an object, and encode boxes with additional regression heads for the scale and offset of a bounding box. Our proposed ConDA builds upon the simple single-stage architecture without relying on default anchors.

2.2 Inherent Interpretability

Although a general definition of interpretability currently doesn't exist, a widespread idea in terms of inherent interpretability is to enforce intra-class concentration and inter-class separation based on distances in the latent space. Hence, two opposing mechanisms are key aspects: (1) clustering positive pairs of the same class and (2) separation of negative samples. The conclusive reasoning process is based on representations that can either be unsupervised prototypes (e.g. ProtoPNet (Chen et al., 2019), ProtoNCE (Li et al., 2021) and CSPP (Feifel et al., 2021b)) or predefined semantic concepts as supervised anchors for the training process (e.g. SupCon (Khosla et al., 2020) and CPD (Feifel et al., 2021a)).

Since semantic concepts or prototypes are represented by feature vectors in the latent space and distance measures are used for predictions, inherent interpretability is closely related to metric learning. Proposed methods are motivated by well-known issues with the *softmax loss* (inner product and cross-entropy loss) arising from the use of the inner product as a similarity measure (Peng and Yu, 2021; Ghiasi-Shirazi, 2019). Loss formulations such as I2CS (Peng and Yu, 2021), SupCon (Khosla et al., 2020) and ProtoNCE (Li et al., 2021) try to enforce intra-class concentration in order to learn more discriminative features.

2.3 Unsupervised Domain Adaptation

Motivated by the easy accessibility of synthetic data and insufficient generalization capabilities of state-of-the-art DNNs, methods for UDA aim at closing the occurring domain gap. Similar to our proposed approach, self-paced learning for object detection (Soviany et al., 2021) uses a teacher-student framework. UMT (Deng et al., 2021) follows a similar approach where a student learns from pseudo labels but they also apply adversarial training for cross-domain distillation. Contrarily, our proposed ConDA does not rely on any kind of style transfer or adversarial training. Regarding UDA for semantic segmentation, ProDA (Zhang et al., 2021) and SePiCo (Xie et al., 2022) leverage prototypes as feature vectors in the latent space to rectify noisy pseudo labels. Softmax-based class predictions are adjusted by distances to prototypes, allowing for more robust classification and generalization in the presence of class imbalances or outliers. Both approaches leverage promising key ideas from the field of inherent interpretability and metric learning. Additionally, DAFormer (Hoyer et al., 2022) specifically designs data augmentation techniques for the target domain.

3 METHODOLOGY

3.1 Generic Pedestrian Detector

In this work, we extend the generic DNN architecture for pedestrian detection (Feifel et al., 2022). We define a PD-DNN $h = f \circ g$ as a composition of the feature extraction f and multiple perception heads g. Figure 1 shows three different PD-DNN architectures that are investigated in our work: (1) PD-DNN for center, scale and offset prediction (CSO), (2) CSO with an auxiliary BPS (CSO+BPS) and (3) CSO with an interpretable reasoning process (IRP) for an auxiliary BPS (CPD). An optional stacked hourglass (HG) (Newell et al., 2016) might increase performance. Fixed deconvolutions (Yu et al., 2018) upsample the predicted BPS.

All PD-DNNs predict a set of 2D bounding boxes for pedestrians. A set of 2D bounding boxes $\mathrm{R} = \{(x_1, y_1, x_2, y_2)\}_{i=1}^{K}$ can be defined as a set of K tuples of four corner coordinates. Predicted bounding

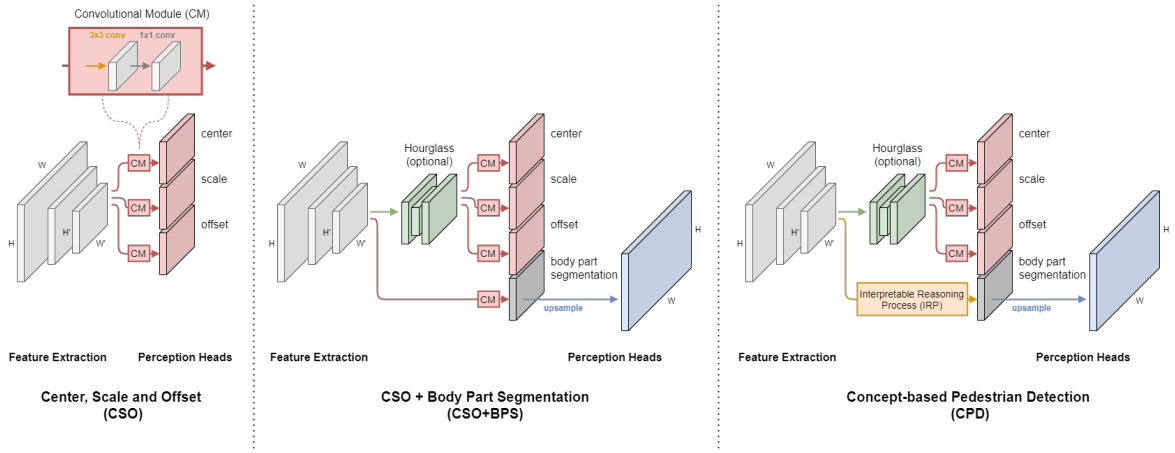

Figure 1: The baseline for our work is given by a simple PD-DNN architecture that applies center, scale and offset heads (CSO) to predict 2D bounding boxes. CSO can be extended with an auxiliary body part segmentation (BPS), denoted as CSO+BPS. The third method uses an interpretable reasoning process (IRP) to predict a body part segmentation for the concept-based pedestrian detection (CPD) (Feifel et al., 2021a).

boxes $\tilde{\mathrm{R}}$ are encoded by the center, scale and offset head. The loss for bounding boxes is defined as $L_{\text{box}} = \lambda_{\text{ce}} L_{\text{ce}} + \lambda_{\text{sc}} L_{\text{sc}} + \lambda_{\text{o}} L_{\text{o}}$. The trade-off weights $\lambda_{\text{ce}}, \lambda_{\text{sc}}$ and λ_{o} are empirically set to $0.01, 1.0$ and 0.1. The scale loss L_{sc} and offset loss L_{o} are defined as smooth L1 loss. Due to the high positive-negative sample discrepancy, L_{ce} uses a focal loss with $\gamma = 4$ (Lin et al., 2017) that is further weighted with a Gaussian map $\mu_{i,j}$ for every pedestrian center (Liu et al., 2019; Feifel et al., 2022). Moreover, an ignore map $o_{i,j}$ highlights ignored bounding boxes. The final center loss is given by

$$L_{\text{ce}} = -\frac{1}{K} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \bar{\mu}_{i,j} (1 - o_{i,j}) (1 - \bar{\nu}_{i,j})^{\gamma} \log \bar{\nu}_{i,j} \quad (1)$$

where

$$\bar{\mu}_{i,j} = \begin{cases} 1 & \text{if } \xi_{i,j} = 1 \\ (1 - \mu_{i,j})^{\beta} & \text{otherwise} \end{cases} \quad (2)$$

$$\bar{\nu}_{i,j} = \begin{cases} \hat{\xi}_{i,j} & \text{if } \xi_{i,j} = 1 \\ 1 - \hat{\xi}_{i,j} & \text{otherwise} \end{cases} \quad (3)$$

Predictions of the center head of the student network h_{θ}^{ce} for a given source image x_S are defined as $\hat{\xi}_{i,j} = h_{\theta}^{\text{ce}}(x_S)$. All binary maps ($\xi_{i,j}, \mu_{i,j}$ and $o_{i,j}$) are created based on a set of ground truth bounding boxes R.

According to Figure 1, the auxiliary BPS can be predicted based on two different transformations: (a) convolutional module (CM) for CSO+BPS or (b) interpretable reasoning process (IRP) for concept-based pedestrian detection (CPD) (Feifel et al., 2021a). We use four predefined body parts and define the set $C_a = \{\text{background}, \text{head}, \text{torso}, \text{arm}, \text{leg}\}$. In the case of the interpretable approach the set of classes is reduced to semantic concepts: $C_b = \{\text{head}, \text{torso}, \text{arm}, \text{leg}\}$. Consequently, the number of classes N_C depends on the chosen transformation: (a) $N_C = |C_a| = 5$ or (b) $N_C = |C_b| = 4$. The perception head for BPS of CSO+BPS or CPD outputs the predictions $\hat{\pi}_{k,i,j}$. Fixed deconvolutions (Yu et al., 2018) compute upsampled confidence scores $\hat{\Pi}_{k,i,j}$: $\pi \in [0,1]^{N_C \times H' \times W'} \rightarrow \Pi \in [0,1]^{N_C \times H \times W}$. Based on one-hot encoded ground truth body part annotations $\Pi_{k,i,j}$, the focal binary cross-entropy loss is formulated as

$$\mathrm{L}_{\text{seg}} = -\frac{1}{N} \sum_{k=1}^{N_C} \sum_{i=1}^{H} \sum_{j=1}^{W} (1 - S_{k,i,j})^{\gamma} \log(S_{k,i,j}) \quad (4)$$

with

$$S_{k,i,j} = \begin{cases} \hat{\Pi}_{k,i,j} & \text{if } \Pi_{k,i,j} = 1 \\ 1 - \hat{\Pi}_{k,i,j} & \text{otherwise} \end{cases} \quad (5)$$

and $N = N_C H W$. The focal term $S_{k,i,j}$ is introduced due to the high class imbalance of negative (background) and positive (body part) pixels.

We define a set of latent representations $Z = \{\mathbf{z}_{i,j}\}_{i,j}^{H',W'}$ with $\mathbf{z} \in [0,1]^{64}$ as an encoding of an image x output by feature extraction $f(x) : \mathcal{X} \rightarrow \mathcal{Z}$. The extracted latent representations are matched with learnable concept vectors $\mathbf{c}_k$ for four concepts (i.e., head, torso, arm and leg) with the help of binary concept masks $\pi_{k,i,j}$. The loss for IRP of CPD is defined as $L_{\text{latent}} = \lambda_{\text{cl}} L_{\text{cl}} + \lambda_{\text{sc}} L_{\text{sep}}^{\text{con}} + \lambda_{\text{sb}} L_{\text{sep}}^{\text{back}}$. We define the mean square loss to learn concept clusters by minimizing the squared euclidean distance of latent representations $\mathbf{z}_{i,j}$ to concept vectors $\mathbf{c}_k$ as

$$L_{\text{cl}} = \frac{1}{M_{\text{cl}}} \sum_{k=1}^{N_C} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \pi_{k,i,j} \|\mathbf{c}_k - \mathbf{z}_{i,j}\|^4 \quad (6)$$

$M_{\text{cl}} = \sum_{k=1}^{N_C} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \pi_{k,i,j}$ describes the number of positive concept pixels. To achieve a high inter-class separation, we introduce a safety margin $\delta'_k = \varepsilon \delta^*_k$ with parameter ε describing a multiple of the critical distance δ^* (Feifel et al., 2021a). The safety margin δ'_k is used to define a weighted loss contribution based on $\psi_k = \exp\left(\frac{\log(z)}{\delta'_k - \delta^*_k} \cdot \left(\delta_k - \delta^*_k\right)\right)$ that penalizes small distances of negative samples. The generic separation loss is defined as

$$L_{\text{sep}} = \frac{1}{M_{\text{sep}}} \sum_{k=1}^{N_C} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \rho_{k,i,j} \left(\psi_k \left(\delta'_k - \|\mathbf{c}_k - \mathbf{z}_{i,j}\|^2\right)\right)^2 \tag{7}$$

with $M_{\text{sep}} = \sum_{k=1}^{N_C} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \rho_{k,i,j}$ as the number of negative pixels. In the case of separation against background ($L_{\text{se}}^{\text{back}}$) or other concepts ($L_{\text{se}}^{\text{con}}$), the one-hot encoded binary mask for negative pixels $\rho_{k,i,j}$ must be adapted accordingly (Feifel et al., 2021a).

Predicted bounding boxes post-processed by the non-maximum suppression (NMS) are a set of detection bounding boxes $\hat{R}$. We only accept bounding boxes $\hat{R}_f = \left\{(x_1, y_1, x_2, y_2) \in \hat{R} \mid \sum_{k=1}^{N_C} \sum_{i=y_1}^{y_2} \sum_{j=x_1}^{x_2} \hat{\Pi}_{k,i,j} > 0\right\}$ that contain pixels of at least one body part.

3.2 Concept-Based Domain Adaptation

We opt for a two-stage methodology for UDA similar to ProDA. Note that, unlike our work, this method focuses on semantic segmentation and not pedestrian detection. Our methodology proposes two consecutive training stages. Initially, a PD-DNN with an auxiliary BPS is trained solely on synthetic data and validated on real data. Hereafter, the pre-trained PD-DNN is used as a starting point for UDA. Since our methodology is related to inherent interpretability and metric learning, we refer to it as concept-based domain adaptation for pedestrian detection (ConDA).

Due to superior generalization, we use CPD as PD-DNN for ConDA. Consequently, ConDA utilizes a structured latent space due to the cluster and separation mechanism of CPD. In contrast to ProDA and SePiCo, we integrate concept vectors (similar to their prototypes) directly into the DNN architecture. Algorithm 1 gives a detailed description of the different training steps for ConDA. The unsupervised domain adaptation starts with a pre-trained CPD h and parameters θ. It was previously trained on images $x_S \in \mathcal{X}_S$ with labels y_S as ground truth from a source domain $\mathcal{X}_S$. In our case, the source domain is represented by synthetic data. We use a conventional self-training approach where pseudo labels are fixed during the second training stage for domain adaptation. Pseudo labels are generated on-the-fly by a pseudo network $\tilde{h}$ with parameters θ for images $x_T \in \mathcal{X}_T$ from the target domain $\mathcal{X}_T$ represented by real data.

Algorithm 1: Pseudocode for ConDA.

Input: Training dataset $(\mathcal{X}_S, \mathcal{Y}_S, \mathcal{X}_T)$, CPD h with pre-trained parameters θ and thresholds $\tau_{\text{ign}}, \tau_{\text{ce}}, \tau_{\text{sep}}, \tau_{\text{cl}}$.
Output: Teacher model $h_{\theta'}$.

Initialize student network h_θ;
Initialize pseudo network: $\tilde{h}_\theta \leftarrow h_\theta$;
Initialize teacher network: $h_{\theta'} \leftarrow h_\theta$;
while $i \leq$ iterations **do**
  Get source images x_S^i and ground truth $\mathcal{Y}_S = (\xi^i, \pi^i, \Pi^i)$;
  Train student h_θ with loss L^S: $L^S = L_{\text{box}} + L_{\text{seg}} + L_{\text{latent}}$;
  Get target images x_T^i;
  Get pseudo targets (ξ'^i, π'^i, Π'^i) from $\tilde{h}_\theta$;
  Train student h_θ with loss L^T: $L^T = L_{\text{box}}^T + L_{\text{seg}}^T + L_{\text{cl}}^T + L_{\text{sep}}^T$;
  Update teacher $h_{\theta'}$:
    - Parameters θ': $\theta'_{i+1} \leftarrow \alpha\theta'_i + (1-\alpha)\theta_i$
    - Running batch norm: $h_{\theta'} = h_{\theta'}(x_T)$;
end

Pseudo bounding boxes are encoded by outputs of the center, scale and offset heads of the pseudo network $\tilde{h}_\theta$. Center predictions for a sample x_T drawn from the target domain are defined as $\tilde{\xi}_{i,j} = \tilde{h}_\theta^{\text{ce}}(x_T)$. Generating pseudo bounding boxes $\tilde{R} = d\left(\left[\tilde{\xi}_{i,j} > \tau_{\text{ce}}\right], \tilde{h}_\theta^{\text{sc}}(x_T), \tilde{h}_\theta^{\text{o}}(x_T)\right)$ is done by applying a decoding function $d()$ (including NMS) with the scale and offset predictions as additional inputs. Applying a center threshold τ_{ce} and the Iverson bracket $[\cdot]$ guarantees that only bounding boxes with high certainty are used as positive pseudo labels. Finally, a hard pseudo target for the center map $\xi'_{i,j}$ can be generated from $\tilde{R}$. To efficiently denoise pseudo labels, ignored (pseudo) bounding boxes are defined as $\tilde{R}_{ign} = d\left(\left[\tilde{\xi}_{i,j} > \tau_{\text{ign}} \wedge \tilde{\xi}_{i,j} \leq \tau_{\text{ce}}\right], \tilde{h}_\theta^{\text{sc}}(x_T), \ldots\right)$. Based on $\tilde{R}_{ign}$ and $\xi'_{i,j}$, pseudo ignore areas $o'_{i,j}$ are defined. To counteract trivial solutions, bounding boxes below a minimal threshold τ_{ign} are seen as negative samples. We empirically set the thresholds τ_{ign} and τ_{ce} to 0.1 and 0.6. The unsupervised center loss for center predictions of student network $\hat{\xi}_{i,j}$ for K predicted pedestrians and hard pseudo targets $\xi'_{i,j}$ for an unlabeled

sample from the target domain is given by

$$L_{\text{ce}}^{T} = -\frac{1}{K}\sum_{i=1}^{H'}\sum_{j=1}^{W'}\bar{\mu}_{i,j}(1-\tilde{o}_{i,j})(1-\bar{\mathbf{v}}_{i,j})^{\gamma}\log\bar{\mathbf{v}}_{i,j} \quad (8)$$

Scale and offset behave accordingly with the same ignore area $\tilde{o}_{i,j}$.

ConDA utilizes the predicted distance-based body part segmentation to strengthen the clustering and further, minimize the distance of positive latent representations to concept vectors. Soft pseudo predictions of the pseudo network for body part segmentation are given by $\tilde{\pi}_{k,i,j}$. We apply the cluster threshold τ_{cl} to define the pseudo concepts masks $\pi'_{k,i,j} = \left[k = \arg\max_{k'} \tilde{\pi}_{k',i,j} \wedge \tilde{\pi}_{k,i,j} > \tau_{\text{cl}}\right]$. We empirically set the threshold τ_{cl} to 0.8. The unsupervised cluster loss with pseudo concept mask $\pi'_{k,i,j}$, concept vectors $\mathbf{c}_k$ and latent representations $\mathbf{z}_{i,j}$ of the student network is defined as

$$L_{\text{cl}}^{T} = \frac{1}{\tilde{M}_{\text{cl}}}\sum_{k=1}^{N_C}\sum_{i=1}^{H'}\sum_{j=1}^{W'}\pi'_{k,i,j}\|\mathbf{c}_k - \mathbf{z}_{i,j}\|^4 \quad (9)$$

where $\tilde{M}_{\text{cl}}$ is the number of pseudo positive concept pixels.

To avoid degenerate solutions, negative latent representations have to be separated. The binary hard pseudo separation mask $\rho'_{k,i,j}$ is based on the soft pseudo targets $\tilde{\pi}_{k,i,j}$ of the pseudo network and can be formulated as $\rho'_{k,i,j} = \left[\tilde{\pi}_{k,i,j} < \tau_{\text{sep}}\right]$. We empirically set the threshold τ_{sep} to 0.1. The final unsupervised separation loss is defined as

$$L_{\text{sep}}^{T} = \frac{1}{\tilde{M}_{\text{sep}}}\sum_{k=1}^{N_C}\sum_{i=1}^{H'}\sum_{j=1}^{W'}\rho'_{k,i,j}\left(\psi_k\left(\delta'_k - \|\mathbf{c}_k - \mathbf{z}_{i,j}\|^2\right)\right)^2 \quad (10)$$

where $\tilde{M}_{\text{sep}}$ is the number of pseudo negative concept pixels.

Finally, pseudo targets for the body part segmentation are defined as $\Pi'_{k,i,j} = \left[k = \arg\max_{k'} \tilde{\Pi}_{k',i,j}\right]$. To efficiently rectify the pseudo labels for negative and positive samples, an ignore area $O'_{k,i,j} = \left[\tau_{\text{sep}} \leq \tilde{\Pi}_{k,i,j} \leq \tau_{\text{cl}}\right]$ is defined. The BPS loss can be formulated as

$$L_{\text{seg}}^{T} = -\frac{1}{\tilde{M}_{seg}}\sum_{k=1}^{N_C}\sum_{i=1}^{H}\sum_{j=1}^{W}(1-O'_{k,i,j}) \\ (1-S_{k,i,j})^{\gamma}\log(S_{k,i,j}) \quad (11)$$

where $\tilde{M}_{\text{seg}} = \sum^{N_C}\sum^{H}\sum^{W}\pi_{k,i,j}$ gives the number of non-ignored pixels for the body part segmentation.

It is well known that DNNs show a tremendous bias towards texture, while shape information is mostly neglected (Geirhos et al., 2018). Limited simulation and rendering resources are the reason that synthetic data is prone to show less variation in shapes (pedestrian posture and perspective) and textures (pedestrian appearance and illumination). To reduce the texture bias and strengthen the shape bias on body parts, we texturize the background and body parts with 50 different textures randomly sampled from the Describable Texture Dataset (DTD) (Cimpoi et al., 2014). Furthermore, we use extensive natural data augmentation techniques to alter brightness, contrast, blurring and other properties. Inspired by (Geirhos et al., 2018; Liu et al., 2019), we propose a shape-enforcing data augmentation strategy (SA) for pedestrian detection.

4 EXPERIMENTS

In our work, we use synthetic data from KI-A, real data from the CityPersons dataset (Zhang et al., 2017) and the Cityscapes-Panoptic-Parts dataset (Meletis et al., 2020) (CS). Results are provided for the validation set of the CityPersons dataset. The domain gap is analyzed based on the performance of the oracle training (CS $\rightarrow$ CS), i.e., supervised training and validation on CS as the target domain. Furthermore, we investigate the following scenarios: Source-only training (KI-A $\rightarrow$ CS), i.e., supervised training on the source domain (KI-A) and validation on the target domain (CS) to analyze generalization and UDA, i.e., supervised training on source domain combined with unsupervised training on target domain and validation of ConDA on target domain (CS).

The most common performance metric for PD-DNNs is the log-average miss rate for the *reasonable* subset (LAMR_{r}) of the CityPersons dataset (Dollar et al., 2011). Another metric is the MR@1FPPI describing the miss rate if we accept one false positive per image (Dollár et al., 2009). To evaluate the segmentation performance of body parts, we utilize the reasonable subset for CityPersons and extend it to the instance segmentation of Cityscapes and further to the body part segmentation given by Cityscapes-Panoptic-Parts and define the mean intersection over union (mIoU_{r}).

4.1 Implementation Details

The following feature extractions are used: MDLA-UP-34 (Yu et al., 2018; Dai et al., 2017), CSP-ResNet-50 (Liu et al., 2019) and FPN-ResNet-50 (Zhang et al., 2020). The Adam optimizer with a learning rate of 1e-4 and a linear warm-up strategy for 2k iterations is applied. With ConDA, the learning rate subsequently decreases linearly. We train for

Table 1: Performance of different PD-DNNs and domain adaptation scenarios. The oracle training (CS → CS) of CSO and CSO+BPS shows competitive performance to state-of-the-art PD-DNNs. Compared to CSO and CSO+BPS, CPD shows superior generalization for source-only training (KI-A → CS). MDLA-UP-34 is used as feature extractor for all methods.

Prediction	Method	Transformation	Hourglass (HG)	CS → CS		KI-A → CS	
				$LAMR_r$	$mIoU_r$	$LAMR_r$	$mIoU_r$
2D BB	CSO	-	-	9.6	-	39.6	-
2D BB + BPS	CSO+BPS	CM	-	9.0	73.4	36.4	44.9
			✓	10.8	70.8	35.6	48.5
	CPD	IRP	-	14.2	59.8	41.0	58.5
			✓	13.3	46.1	**34.0**	**62.1**

Table 2: Performance for the CityPersons validation dataset considering different domain adaptation scenarios. We can see that ConDA substantially exceeds the baseline performance (comparing underlined values) by benefiting from CPD. The MR@1FPPI is shown for different subsets (i.e., reasonable (R), bare (B) and large (L)). MDLA-UP-34 is used as feature extractor for all methods.

Method	Scenario	$LAMR_r$	MR@1FPPI			$mIoU_r$	IoU_r			
			R	B	L		Head	Torso	Arm	Leg
CSO	CS → CS	9.6	3.9	5.5	5.5	-	-	-	-	-
CSO+BPS		9.0	3.9	2.2	2.3	73.4	75.2	64.6	53.0	74.7
CPD		14.2	4.9	3.3	3.4	59.8	0.0	66.6	63.5	70.0
CSO	KI-A → CS	39.6	23.7	18.5	14.6	-	-	-	-	-
CSO+BPS		<u>35.6</u>	21.7	14.3	12.3	<u>48.5</u>	46.4	19.1	20.3	57.9
CPD		34.0	21.5	14.2	12.5	62.1	54.6	51.6	43.8	60.9
CPD (w/ SA+WU)		28.7	13.1	7.5	6.8	62.5	55.2	44.9	48.5	64.2
ConDA	UDA	<u>23.0</u>	10.0	5.1	6.0	<u>65.8</u>	57.9	56.0	51.2	64.0

a maximum of 50k iterations on 2 GPUs with a batch size of 8. The parameters θ' of the teacher network are constantly averaged based on a student-teacher framework (Tarvainen and Valpola, 2017). For inference, only center points with a confidence score > 0.1 and bounding boxes with height ≥ 50 pixels are considered. For post-processing, we apply a Greedy-NMS with a threshold of 0.5.

4.2 Results

Since ConDA is a two-stage approach, we first analyze the generalization capabilities of source-only training (KI-A → CS) regarding the $LAMR_r$ and $mIoU_r$ of different PD-DNNs in Table 1. The naive source-only training is seen as our baseline for all domain adaptation strategies of ConDA. We can show that CSO+BPS (w/o HG) improves the $LAMR_r$ performance by 0.6% absolute for the oracle scenario compared to CSO. The performance of source-only training for CSO+BPS (w/o HG) decreases absolutely by 27.4% $LAMR_r$ and 28.5% $mIoU_r$ compared to oracle training. CPD (w/ HG) shows the best generalization with 34.0% $LAMR_r$ and 62.1% $mIoU_r$. It outperforms CSO+BPS (w/ HG) by a large absolute margin of 1.6% $LAMR_r$ and 13.6% $mIoU_r$. The additional stacked hourglass (HG) is particularly useful for CPD. Due to better generalization compared to CSO and CSO+BPS, CPD is used for our proposed ConDA approach.

Table 3: Ablation study to validate the generalization of source-only training for CPD (KI-A → CS), different training strategies and feature extractions.

Feat. Extr.	SA	WU	$LAMR_r$	$mIoU_r$
MDLA-UP-34	-	-	34.0	62.1
MDLA-UP-34	✓	-	32.0	54.8
MDLA-UP-34	-	✓	39.4	57.5
MDLA-UP-34	✓	✓	**28.7**	**62.5**
FPN-ResNet-50	✓	✓	35.7	56.3
CSP-ResNet-50	✓	✓	33.0	57.4

Results of the ablation study for different training strategies are shown in Table 3 and are complemented by the performance of two comparable feature extractions. Improving the initial performance of CPD seems reasonable since it serves as a starting point for

Figure 2: Inference results: original image (top row), ground truth bounding boxes in blue, ignored bounding boxes in red and ground truth body part segmentation (middle row) and ConDA predictions (bottom row).

ConDA and has a great impact on the final performance. It can be shown that SA in combination with a linear warm-up strategy for the learning rate (WU) offers the best performance in terms of $LAMR_r$ and $mIoU_r$ for MDLA-UP-34. However, Table 3 shows an absolute $LAMR_r$ increase of 5.4% for WU, demonstrating strong overfitting to the source domain. WU can only contribute in combination with SA, which means extensive data augmentation. CSP-ResNet-50 and FPN-ResNet-50 cannot benefit from the proposed strategies, thus the following experiments are only performed with MDLA-UP-34. SA and WU are also applied to the second training stage of ConDA. Figure 2 shows exemplary inference results of ConDA for the CityPersons validation dataset.

Compared to the naive source-only training, Table 2 shows that ConDA achieves an absolute improvement of 16.6% $LAMR_r$ compared to CSO, stating the successful UDA towards the real CityPersons dataset. We also show that ConDA improves segmentation performance for often overlapping and thus difficult body parts such as torso and arm. ConDA substantially reduces MR@1FPPI compared to the naive source-only training and misses the oracle performance by only 6.1% absolute for the reasonable subset. As to be expected, the performance gap is substantially lower for easier subsets (i.e, bare and large). We see an absolute MR@1FPPI of only 5.1% and 6% respectively. Hence, ConDA nearly matches the performance of state-of-the-art PD-DNNs. Due to the applied metric learning, CPD has well-separated clusters of discriminative features for body parts leading to better generalization.

5 CONCLUSION

In our work, we propose ConDA as a novel method for domain adaptation for pedestrian detection that enhances supervised learning from synthetic data with unsupervised learning from real and unlabeled data. We show that enforcing intra-class concentration of semantic concepts for pedestrians and inter-class separation leads to a better generalization. Compared to naive training on only synthetic data, ConDA substantially increases the performance of pedestrian detection and an auxiliary body part segmentation by a large margin on real data. In conclusion, our proposed method ConDA can be seen as a promising step towards using low-cost synthetic data through domain adaptation for pedestrian detection based on semantic concepts.

ACKNOWLEDGEMENTS

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project “Methoden und Maßnahmen zur Absicherung von KI basierten Wahrnehmungsfunktionen für das automatisierte Fahren (KI-Absicherung)”. The authors would like to thank the consortium for the successful cooperation.

REFERENCES

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. (2014). Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. (2017). Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773.

Deng, J., Li, W., Chen, Y., and Duan, L. (2021). Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2009). Pedestrian detection: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 304–311.

Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2011). Pedestrian detection: An evaluation of the state of the

art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761.

Feifel, P., Bonarens, F., and Köster, F. (2021a). Leveraging interpretability: Concept-based pedestrian detection with deep neural networks. In *Computer Science in Cars Symposium*, CSCS '21. Association for Computing Machinery.

Feifel, P., Bonarens, F., and Koster, F. (2021b). Reevaluating the safety impact of inherent interpretability on deep neural networks for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29–37.

Feifel, P., Franke, B., Raulf, A., Schwenker, F., Bonarens, F., and Köster, F. (2022). Revisiting the evaluation of deep neural networks for pedestrian detection. In *Proceedings of the Workshop on Artificial Intelligence Safety 2022*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Ghiasi-Shirazi, K. (2019). Generalizing the convolution operator in convolutional neural networks. *Neural Processing Letters*, 50(3):2627–2646.

Haselhoff, A., Kronenberger, J., Kuppers, F., and Schneider, J. (2021). Towards black-box explainability with gaussian discriminant knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21–28.

Hoyer, L., Dai, D., and Van Gool, L. (2022). Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935.

Khan, A. H., Munir, M., van Elst, L., and Dengel, A. (2022). F2dnet: Fast focal detection network for pedestrian detection. *arXiv preprint arXiv:2203.02331*.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. (2020). Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR.

Li, J., Liao, S., Jiang, H., and Shao, L. (2020). Box Guided Convolution for Pedestrian Detection. In *28th ACM International Conference on Multimedia*, pages 1615–1624.

Li, J., Zhou, P., Xiong, C., and Hoi, S. C. H. (2021). Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations ICLR*.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal Loss for Dense Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2980–2988.

Liu, W., Liao, S., Ren, W., Hu, W., and Yu, Y. (2019). High-level semantic feature detection: A new perspective for pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5187–5196.

Meletis, P., Wen, X., Lu, C., de Geus, D., and Dubbelman, G. (2020). Cityscapes-panoptic-parts and pascal-panoptic-parts datasets for scene understanding. *arXiv preprint arXiv:2004.07944*.

Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer.

Peng, H. and Yu, S. (2021). Beyond softmax loss: Intra-concentration and inter-separability loss for classification. *Neurocomputing*, 438:155–164.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Soviany, P., Ionescu, R. T., Rota, P., and Sebe, N. (2021). Curriculum self-paced learning for cross-domain object detection. *Computer Vision and Image Understanding*, 204:103166.

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Xie, B., Li, S., Li, M., Liu, C. H., Huang, G., and Wang, G. (2022). Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *arXiv preprint arXiv:2204.08808*.

Yu, F., Wang, D., Shelhamer, E., and Darrell, T. (2018). Deep layer aggregation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2403–2412.

Zhang, J., Lin, L., Zhu, J., Li, Y., Chen, Y.-c., Hu, Y., and Hoi, C. S. (2020). Attribute-aware pedestrian detection in a crowd. *IEEE Transactions on Multimedia*.

Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., and Wen, F. (2021). Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424.

Zhang, S., Benenson, R., and Schiele, B. (2017). Citypersons: A diverse dataset for pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3221.