

Learning Cascaded Detection Tasks with Weakly-Supervised Domain Adaptation

Niklas Hanselmann^{1,2}, Nick Schneider¹, Benedikt Ortelt³ and Andreas Geiger^{2,4}

Abstract—In order to handle the challenges of autonomous driving, deep learning has proven to be crucial in tackling increasingly complex tasks, such as 3D detection or instance segmentation. State-of-the-art approaches for image-based detection tasks tackle this complexity by operating in a cascaded fashion: they first extract a 2D bounding box based on which additional attributes, e.g. instance masks, are inferred. While these methods perform well, a key challenge remains the lack of accurate and cheap annotations for the growing variety of tasks. Synthetic data presents a promising solution but, despite the effort in domain adaptation research, the gap between synthetic and real data remains an open problem. In this work, we propose a weakly supervised domain adaptation setting which exploits the structure of cascaded detection tasks. In particular, we learn to infer the attributes solely from the source domain while leveraging 2D bounding boxes as weak labels in both domains to explain the domain shift. We further encourage domain-invariant features through class-wise feature alignment using ground-truth class information, which is not available in the unsupervised setting. As our experiments demonstrate, the approach is competitive with fully supervised settings while outperforming unsupervised adaptation approaches by a large margin.

I. INTRODUCTION

Modern deep learning-based vision systems have made remarkable strides in recent years, enabled in no small part by leveraging large amounts of labeled training data. Unfortunately, acquiring annotated data is often expensive, time consuming and, for some tasks, also inaccurate or nearly impossible. To alleviate this problem, synthetic data obtained from game-engines or purpose-built simulation environments might be an alternative, promising a large amount of training data with highly accurate, inexpensive annotations. However, despite ever-improving photorealism, current state-of-the-art synthetic datasets still fail to resemble actual real world sensor data. Subsequently, machine learning models which are trained on a synthetic dataset typically perform poorly if applied to real world data. In order to bridge this domain gap, several works on domain adaptation have been proposed. Most commonly, they assume an *unsupervised* regime, in which we have access to fully-labeled data in a source domain and only unlabeled data in a target domain. This setting is particularly difficult and while the community continues to

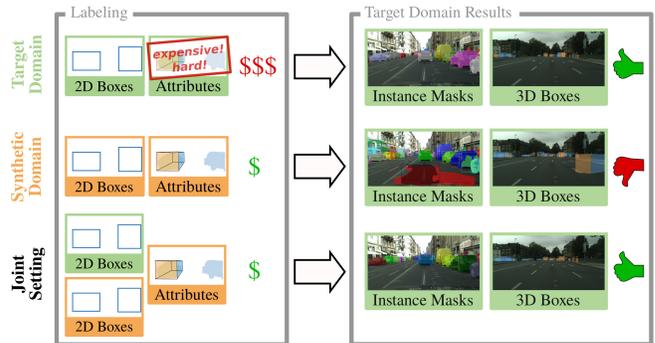


Fig. 1: **Proposed Setting.** We propose a weakly-supervised domain-adaptation setting that enables learning cascaded detection tasks at a reduced annotation effort while still achieving competitive performance by utilizing weak 2D bounding box labels in both domains.

make progress in tackling the open problems in unsupervised domain adaptation, the gap in performance when compared to target domain supervision remains significant, limiting the utility of these models for practical applications, such as autonomous driving. Is there an alternative which allows for accurate results without the need of expensive labeling in the target domain?

In this paper, we propose a setting which exploits the decomposability of complex detection tasks into a cascaded structure. This concept is central to many image-based detection approaches where objects are first detected and, based on the detection, additional attributes such as instance masks [1] or 2D/3D poses [2] are inferred. Likewise, labeling can now also be decomposed into 2D bounding box labels, which are inexpensive to annotate and for which annotated datasets already exist in many domains, and additional, more complex attribute labels, which are often expensive or difficult to obtain with sufficient accuracy. Motivated by this, we investigate the use of weak supervision in the form of 2D bounding box annotations in the target domain in conjunction with full supervision in the source domain. Our key hypothesis is that a large part of the domain shift might already be explained in the detection stage, while the following cascaded stage transfers more readily given accurate detections. To test this hypothesis, we consider two common cascaded detection tasks in this paper: Instance segmentation and monocular 3D detection.

Although we find that a strategy of jointly training on both domains with the available weak supervision can already work surprisingly well, there is no guarantee that the learned

*This publication was created as part of the research project "KI Delta Learning" (project number: 19A19013A) funded by the Federal Ministry for Economic Affairs and Energy (BMWi) on the basis of a decision by the German Bundestag.

¹Mercedes-Benz AG, R&D, Stuttgart, Germany

²University of Tübingen, Tübingen, Germany

³Robert Bosch GmbH, Stuttgart, Germany

⁴Max Planck Institute for Intelligent Systems, Tübingen, Germany

Primary contact: niklas.hanselmann@daimler.com

feature representation will always be useful for the cascaded stage in the target domain. As evidenced by our experiments in Sec. IV, joint training can in fact lead to domain-specific features, rendering the cascaded stage less transferable. Borrowing from the unsupervised domain adaptation literature, we address this problem by encouraging domain-invariant representations. A typical approach is to align marginal feature distributions across domains. However, this does not guarantee a low target domain error for the specific task and can even be detrimental if the label distributions differ across domains [3]. Recent work has addressed this problem by considering class information during feature alignment [4], [5], aiming at invariance of the conditional feature distributions given the corresponding labels. However, doing so in the absence of ground-truth target domain labels is challenging due to error accumulation [6]. Our setting allows us to sidestep this issue as it enables us to accurately condition the feature distribution alignment on the ground-truth class at instance-level.

In summary, we make the following contributions:

- We propose and systematically analyze a weakly supervised domain adaptation setting for cascaded detection tasks, where 2D bounding box annotations are available in both domains.
- We show that in this setting, models adapted using existing techniques can now be competitive with models that are fully supervised in the target domain at a significantly reduced annotation effort. Furthermore, they outperform unsupervised adaptation approaches by a large margin, justifying the additional supervision.
- We analyze the role of weak supervision in bridging the domain shift both in isolation and in conjunction with feature distribution alignment.

With this work we hope to both inspire and provide a baseline for future research on domain adaptation leveraging weak labels. Supplementary material is available at: <https://lasnik.github.io/wdsa/>.

II. RELATED WORK

The domain adaptation problem has received great interest in recent years, especially in the unsupervised setting, with several studies in the context of diverse tasks such as classification [7], [4], semantic segmentation [8], [9], [10] and more recently object detection [11], [12], [13]. In the following, we review the most related work on unsupervised domain adaptation in general and in the context of object detection specifically, as well as cross-domain learning with weak supervision.

Unsupervised Domain Adaptation: Underpinned by the theory proposed in [14], the majority of approaches in the unsupervised regime tackle domain adaptation by attempting to mitigate the discrepancy between data distributions across domains. Building on the progress in generative modeling and image-to-image translation [15], [16], one line of work seeks to achieve this directly in image space [17], [8], [9], aiming at distributional alignment at *pixel-level*. A popular

alternative has been to instead focus on alignment at *feature-level*. Methods in this category typically attempt to match the marginal distributions of source and target domain features. This is done either explicitly by minimizing a divergence measure [18], [19] or implicitly through domain-adversarial training [7], [4]. Orthogonal to these approaches, another line of work leverages unlabeled auxiliary tasks and utilizes them in a self-training scheme alongside the main task [17], [20], [21]. Many of these works introduced concepts that are now central to the domain adaptation literature, but focus mostly on image classification and semantic segmentation, while we are interested in detection-based tasks.

Cross-Domain Object Detection: In [11], Chen *et al.* follow the domain-adversarial training paradigm in a first study of unsupervised domain-adversarial adaptation in the context of object detection: They propose to align features at both image- and instance-level by means of domain-adversarial training. Motivated by the hypothesis that transferability of features might decrease towards deeper layers, Saito *et al.* [12] employ strong and weak alignment strategies at image-level for shallow and deep features, respectively. Several works follow a similar notion by proposing different hierarchical alignment schemes [22], [23], [24], [25], while others focus on mining descriptive region- [26] or instance-level [13] features for alignment. To ensure discriminativity of the learned representation, recent work [24], [13] proposes aligning category prototype embeddings by relying on the detection model’s class estimates in the target domain. In [27], a few-shot adaptive cross-domain detection setting is studied, where a small number of annotated target domain images are available, for which the per-class feature distributions are aligned with ground-truth class labels. In a slight departure from the prevalent paradigm of learning domain-invariant representations, [28], [29] view the cross-domain detection problem as one of robust learning from self-generated noisy pseudo-labels. Although this body of work marks extensive progress, current unsupervised domain adaptation methods still fail to close the domain gap - often by a large margin - rendering them of limited use for practical applications.

Cross-Domain Learning with Weak Supervision: While the unsupervised regime has been extensively studied, domain adaptation with access to weak annotations from an auxiliary task in the target domain has received comparatively less attention. Previous works utilize depth images as weak supervision to aid in transferring a semantic segmentation model. As the relationship between depth and semantics is not trivial, these works focus mainly on strategies to transfer knowledge both across domains as well as between the auxiliary task and the task of interest, either by learning direct mapping functions [30] or through knowledge distillation [31]. Although an intriguing direction, like unsupervised approaches, these early works are unsuccessful in closing the domain gap. In [32], Wang *et al.* also consider the task of semantic segmentation, but leverage 2D bounding boxes as weak supervision in both domains. They consider a multitask detection and semantic segmentation architecture,

where both tasks share a single backbone network and apply class-agnostic image-level and class-wise instance-level adversarial feature alignment. Here again, the exact relationship between detection and semantic segmentation is not immediately obvious, yielding results comparable to unsupervised domain adaptation methods. In contrast to these works, we specifically consider a cascaded setting where the auxiliary task is *subsumed* by the task of interest and its potential benefit is more readily apparent. Similarly to ours, another line of work follows this same notion, for example by leveraging attribute annotations for fine-grained recognition [33], image-level class annotations for object detection [34] or 2D pose annotations for 3D human pose estimation [35]. These works consider a specific task and rely on explicit constraints induced by its relationship to the auxiliary task, such as consistency between attribute-level and fine-grained predictions, the ability to generate pseudo-labels from weak annotations or geometric correspondences between poses in 2D and 3D. In contrast, we are interested in leveraging 2D bounding box annotations for an entire class of related problems, namely multiple cascaded detection tasks, without focusing on explicit, task-specific regularization.

III. METHOD

In this section, we present our weakly-supervised domain adaptation approach. First, we formally describe the learning problem resulting from our novel setting in Sec. III-A. In order to solve this problem, we use the cascaded framework described in Sec. III-B, which allows to leverage weak annotations in the target domain. Finally, in Sec. III-C and III-D, we describe strategies to optimize the cascaded detection model in our setting with the aim of robust and accurate target domain performance.

A. Learning Problem

Image-based object detection aims to extract the image location, typically in the form of 2D bounding box coordinates $\mathbf{b} \in \mathcal{B} \subset \mathbb{R}^{k \times 4}$, and semantic class information $\mathbf{c} \in \mathcal{C} = \{c_0, c_1, \dots, c_{L-1}\}^k$ of all k objects in the image, where L is the number of classes under consideration. There exist various extensions to the task that additionally estimate more complex attributes \mathbf{a} for each object, such as precise shape or 3D information. However, these additional attributes are often expensive or difficult to label and may therefore not be available in the target domain \mathcal{T} . A natural solution is to use training data from another domain \mathcal{S} , which contains the missing labels.

Let \mathcal{D}^s and \mathcal{D}^t denote the source and target domain datasets respectively, with $\mathcal{D}^s = \{(\mathbf{x}_i^s, \mathbf{b}_i^s, \mathbf{c}_i^s, \mathbf{a}_i^s)\}_{i=1}^{N_s}$, where \mathbf{x}_i^s is the input image and N_s is the number of samples in the dataset. The notation follows analogously for the target domain $\mathcal{D}^t = \{(\mathbf{x}_i^t, \mathbf{b}_i^t, \mathbf{c}_i^t)\}_{i=1}^{N_t}$, where we have access to weak supervision in the form of 2D bounding box coordinates \mathbf{b} and class information \mathbf{c} , but no supervision for the additional attributes. We assume that both domains share a common label space, but follow different data distributions. Given \mathcal{D}^s and \mathcal{D}^t , our goal is then to learn a cascaded

detection model with source domain supervision on the attributes \mathbf{a} that will transfer well to the target domain, while utilizing the supervision on \mathbf{b} and \mathbf{c} that is available in both domains.

While the approach is applicable to any cascaded detection task, we focus on the problems of instance segmentation and monocular 3D detection as examples in this work. Consequently, the additional attributes \mathbf{a} represent pixel-wise instance masks and 3D shape and pose for instance segmentation and 3D detection, respectively.

B. Cascaded Detection Model

For the cascaded detection framework, we follow the state of the art [1], [36] and choose a multi-stage architecture, a design choice that also follows naturally from our setting, as it explicitly encodes the decomposition of the cascaded detection task. We hence have two main components, (1) a base network responsible for 2D detection and (2) an attribute network that builds on the detection network’s predictions to estimate additional attributes for each detection. In the following we use the hat operator to denote model predictions.

Detection network: The detection network G_θ is a convolutional neural network with parameters θ , that maps the input image $\mathbf{x} \in \mathcal{X}$ to estimates of the per-object 2D bounding boxes $\hat{\mathbf{b}} \in \mathcal{B}$ and semantic class labels $\hat{\mathbf{c}} \in \mathcal{C}$:

$$G_\theta(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{B} \times \mathcal{C} \times \mathcal{Z} \quad (1)$$

It consists of a backbone, which extracts an image-level feature representation from the input, and a detection module, that predicts $\hat{\mathbf{b}}$ and $\hat{\mathbf{c}}$ from that representation. In addition to 2D detections, this network also provides a per-object feature representation $\hat{\mathbf{z}} \in \mathcal{Z}$ for the attribute network by applying ROIAlign [1] on the image-level representation using the estimated bounding box coordinates $\hat{\mathbf{b}}$.

Attribute network: The attribute network A_ω is a neural network with parameters ω that, given detections described by $\hat{\mathbf{b}}$ and $\hat{\mathbf{c}}$, estimates the additional attributes $\hat{\mathbf{a}} \in \mathcal{A}$ from their object-level feature representation $\hat{\mathbf{z}}$:

$$A_\omega(\hat{\mathbf{z}}, \hat{\mathbf{b}}, \hat{\mathbf{c}}) : \mathcal{B} \times \mathcal{C} \times \mathcal{Z} \rightarrow \mathcal{A} \quad (2)$$

The specific instantiation of both \mathbf{a} and consequently A_ω depends on the cascaded detection task under consideration.

C. Joint Training with Weak Supervision

Given this framework, how do we best utilize \mathcal{D}^s and \mathcal{D}^t to learn the optimal model parameters θ^* and ω^* , such that the resulting model will be robust and accurate in the target domain? Let us begin by considering the unsupervised domain adaptation setting, where we have to rely solely on source domain supervision, resulting in the following combined loss function for both the detection and the attribute network:

$$\mathcal{L}_{source} = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathcal{L}_{det}(G_\theta(\mathbf{x}_i^s), \mathbf{b}_i^s, \mathbf{c}_i^s) + \mathcal{L}_{att}(A_\omega(G_\theta(\mathbf{x}_i^s)), \mathbf{a}_i^s) \quad (3)$$

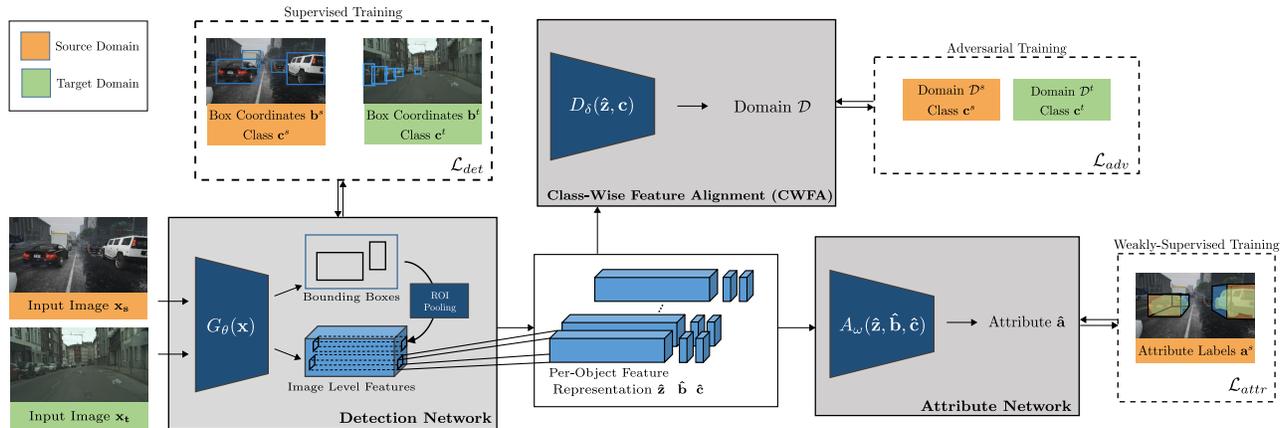


Fig. 2: **Overview of the presented approach.** The goal of our approach is to detect objects in a target domain image by describing them in terms of their location \mathbf{b} and class \mathbf{c} as well as additional, more complex attributes \mathbf{a} , such as precise shape or 3D information. We consider a weakly supervised domain adaptation setting, where labels for the attributes are only available in the source domain (e.g. synthetic data), while easier-to-obtain 2D bounding boxes are available in both domains. We utilize a cascaded detection framework where first a detection network G_θ , trained with supervision on both domains, extracts $\hat{\mathbf{b}}$ and $\hat{\mathbf{c}}$ as well as per-object feature representations $\hat{\mathbf{z}}$. Based on the detection network’s outputs the attribute network A_ω , supervised on the source domain only, then infers $\hat{\mathbf{a}}$. To encourage representations that are meaningful in both domains, the distributions over features $\hat{\mathbf{z}}$ are aligned via a class-conditional domain classifier D_δ .

The exact formulation of \mathcal{L}_{det} and \mathcal{L}_{att} depends on the specific instantiations of G_θ and A_ω , which are described in Sec. IV-A. In our setting, we also assume access to target domain supervision for the detection network:

$$\mathcal{L}_{target} = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathcal{L}_{det}(G_\theta(\mathbf{x}_i^t), \mathbf{b}_i^t, \mathbf{c}_i^t) \quad (4)$$

We denote the sum of both loss terms as \mathcal{L}_{joint} , yielding the following combined objective:

$$(\theta^*, \omega^*) = \underset{\theta, \omega}{\operatorname{argmin}} \mathcal{L}_{joint} \quad (5)$$

where the detection network is trained jointly with supervision from both domains. A key hypothesis in this work is that such a joint training strategy can potentially aid the transferability of the cascaded detection model across domains in two main aspects: Firstly, since any prediction of the attributes $\hat{\mathbf{a}}$ implicitly assumes a valid and accurate 2D detection, the performance of the detection network is crucial. Training it jointly on both domains ensures that this assumption is less likely to be broken, decoupling the adaptation efforts for the attribute network from the 2D detection task. Secondly, the detection network might learn shared representations across domains when trained jointly, rendering any attribute network making predictions based on those representations more readily transferable.

D. Class-Wise Feature Alignment

Although we find that this joint training strategy works well in some scenarios, it may not always be optimal. Firstly, its success depends highly on the degree to which the attribute network reuses the feature representations that emerge by training the detection network jointly. If both

networks build on distinct, task-specific features in the backbone such a strategy cannot hope to improve cross domain performance beyond increased accuracy of the detection network. Secondly, there is no guarantee that joint training always results in a shared representation across domains. Should the representations indeed be domain-specific, the attribute network might not yield the desired results in the target domain.

To counteract this, we follow the same paradigm of regularizing the learned representation to be domain-invariant that is central to many state-of-the-art unsupervised cross-domain detection methods. Since in our setting ground-truth information on the object coordinates and semantic class is available in the target domain, we can perform accurate object-level class-wise feature alignment (CWFA) without error accumulation. To this end, we utilize a class-conditional domain classifier D_δ , which tries to separate the features according to their domain given their ground-truth class. At the same time the detection network G_θ is optimized to output features that maximally confuse the domain classifier in an adversarial fashion. More concretely, D_δ maps the object-level features $\hat{\mathbf{z}}$ to a probability of belonging to the target domain for each class and then outputs the probability associated with the current ground-truth class \mathbf{c} :

$$D_\delta(\hat{\mathbf{z}}, \mathbf{c}) : \mathcal{Z} \times \mathcal{C} \rightarrow [0, 1] \quad (6)$$

This approach of considering class information is similar to previous work in the context of few-shot cross-domain detection [27] and is particularly effective in our proposed setting, as we have the ground-truth class available for every example. Following [12], we use a focal loss [37] with

weighting parameter γ to optimize the domain classifier:

$$\begin{aligned} \mathcal{L}_{adv} = & \\ & - \frac{1}{N_s} \sum_{i=1}^{N_s} (1 - D_\delta(G_\theta(\mathbf{x}_i^s), \mathbf{c}_i^s))^\gamma \log D_\delta(G_\theta(\mathbf{x}_i^s), \mathbf{c}_i^s) \\ & - \frac{1}{N_t} \sum_{i=1}^{N_t} D_\delta(G_\theta(\mathbf{x}_i^t), \mathbf{c}_i^t)^\gamma \log (1 - D_\delta(G_\theta(\mathbf{x}_i^t), \mathbf{c}_i^t)) \end{aligned} \quad (7)$$

We thus obtain the following overall objective:

$$(\theta^*, \omega^*, \delta^*) = \underset{\theta, \omega}{\operatorname{argmin}} \underset{\delta}{\operatorname{argmax}} \mathcal{L}_{joint} - \lambda_{adv} \mathcal{L}_{adv} \quad (8)$$

where G_θ and A_ω are optimized to minimize both the 2D- and attribute detection losses according to the joint training strategy while simultaneously maximally confusing the domain classifier D_δ , which is optimized for low domain classification error, and λ_{adv} is a trade-off parameter. As in previous works [11], [12], [24], the adversarial min-max game is realized through the use of a Gradient Reversal Layer (GRL) [7] at the input of the domain classifier, which flips the sign of the gradients and scales them by a weighting parameter λ_{grl} during the backward pass.

IV. EXPERIMENTS

In the following, we present our experimental results on various synthetic and real datasets. To this end, we first describe our experimental setup in Sec. IV-A and subsequently compare our method to several baseline approaches on the tasks of instance segmentation and monocular 3D detection in Sec. IV-B. Finally, we provide an ablation study in Sec. IV-C.

A. Experiment Setup

Datasets: We perform experiments on four datasets: (1) *Cityscapes (CS)* [38] is a scene understanding benchmark containing 5k diverse real-world urban driving scenes. (2) *Foggy Cityscapes (FCS)* [39] is an extension to CS, that adds synthetic fog to the original scenes. As a result, both have exactly the same label distribution. (3) *Synscapes (SYN)* [40] is a collection of 25k photorealistic synthetic urban traffic scenes from a custom rendering engine. (4) *VIPER (VIP)* [41] contains over 100k synthetic images obtained from the video game *Grand Theft Auto V*.

With these datasets we build three adaptation scenarios: SYN \rightarrow CS, VIP \rightarrow CS and CS \rightarrow FCS. For each scenario, we create a common label space between both domains. Where possible, we correct differences in labeling policies during preprocessing, otherwise removing incompatible classes from consideration.

Evaluation Protocol: We report the performance on the validation split of the target domain dataset using the official evaluation protocols [38], [42]. We use the mean average precision (mAP) and mean detection score (mDS) as metrics for instance segmentation and monocular 3D detection, respectively. We run each experiment three times using

Method	Target Supervision		SYN \rightarrow	CS \rightarrow	VIP \rightarrow
	2D	Mask	CS	FCS	CS
Source Only			17.0	13.9	6.9
DAFRCNN [11]			17.5	23.3	9.1
SWDA [12]			17.5	22.1	6.7
WSJT (Ours)	✓		32.1	30.3	30.3
+ CWFA	✓		31.3	30.3	30.2
Oracle	✓	✓	33.6	30.2	36.1

TABLE I: **Results on instance segmentation.** We report the mean average precision.

Method	Target Supervision		SYN \rightarrow	CS \rightarrow
	2D	3D	CS	FCS
Source Only			14.2	13.7
DAFRCNN [11]			15.4	19.5
SWDA [12]			14.8	17.8
WSJT (Ours)	✓		20.4	22.6
+ CAFA	✓		21.1	24.1
+ CWFA	✓		23.4	24.2
Oracle	✓	✓	25.0	24.7

TABLE II: **Results on monocular 3D detection.** Values are mean Detection Scores [42].

different random seeds and report the performance of the best run at the last training iteration.

Baselines & Comparison: As a lower bound we train a model on the source domain only and apply it on the target domain during testing. As an upper bound, we train an oracle model with full target domain supervision. In addition, we also consider the unsupervised cross-domain detection approaches of Chen et al. [11] and Saito et al. [12]. We re-implemented their work in order to extend it to our datasets and setting. We compare these baselines against models adapted in our proposed setting using both just the weakly-supervised joint training strategy in isolation (*WSJT*) and in conjunction with class-wise feature alignment (*WSJT + CWFA*). On the task of monocular 3D detection, we additionally report results of the latter configuration with *class-agnostic* feature alignment (*WSJT + CAFA*), for which we use the instance-level alignment module in [11], to evaluate the benefit of considering class information. To the best of our knowledge, there is no prior work on leveraging 2D bounding box annotations for domain adaptation in the context of instance segmentation and monocular 3D detection, thus we provide an additional ablation study of the proposed weakly supervised setting in Sec. IV-C.

Implementation and Training Details: As the multi-stage cascaded detection framework we use a standard implementation¹ of Mask-RCNN [1] with a Resnet-50 [43] backbone in conjunction with a Feature Pyramid Network (FPN) [44]. For monocular 3D detection, we replace the mask head with a reimplementation of the 3D bounding box regressor proposed in [36], where we omit the confidence rescaling

¹<https://github.com/facebookresearch/detectron2>

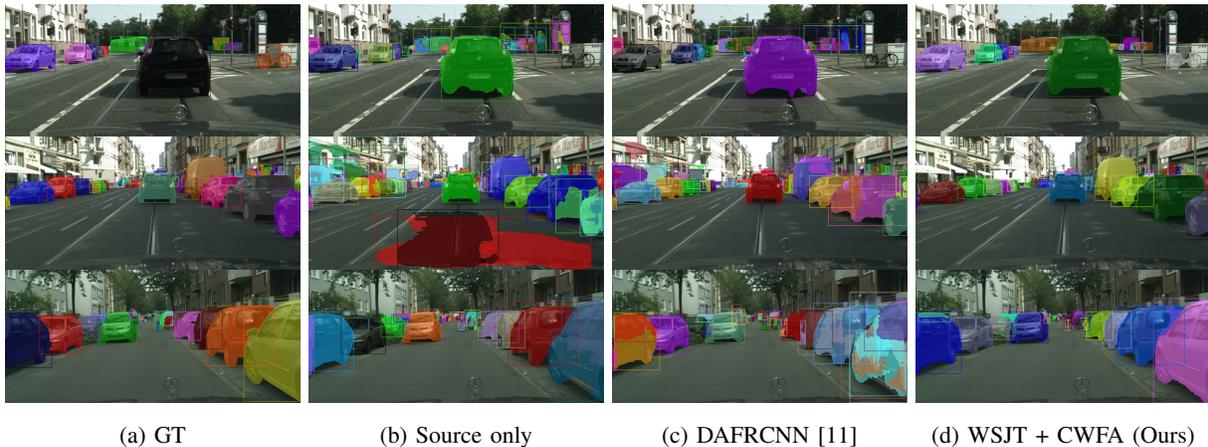


Fig. 3: **Instance segmentation results (SYN \rightarrow CS)**. In contrast to our approach, both baselines exhibit a higher false positive rate. Besides improved detection performance, also the contours of our segmentations are typically more accurate.

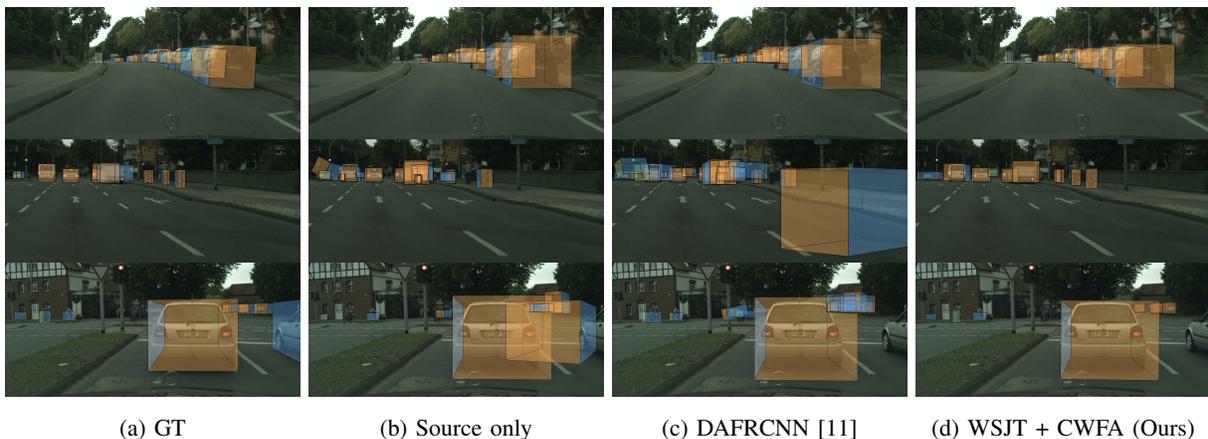


Fig. 4: **Monocular 3D detection results (SYN \rightarrow CS)**. Besides less false positive detections, our approach also yields favorable orientation estimates in comparison to the baseline methods.

Method	Instance Segmentation (mAP)	3D Detection (mDS)
Source Only	27.7	22.8
DARCNN [11]	27.4	23.7
SWDA [12]	27.2	24.5
WSJT (Ours)	36.2	23.9
+ CWFA	34.1	31.6
Oracle	45.5	34.5

TABLE III: **Ablation Study**. To evaluate the transferability of the learned feature representations, we decouple the final performance from that of the 2D detection network, by using ground-truth 2D bounding boxes at test-time (SYN \rightarrow CS).

branch. This regressor works on amodal 2D bounding boxes which, in contrast to other datasets [45], [46], are currently not available in Cityscapes. We therefore compute them as projections of the 3D bounding boxes onto the image plane. For the class-conditional domain classifier we reuse the instance-level architecture of Chen et al. [11] and extend the last fully-connected layer to output L per-class domain

probabilities. We adapt the methods of Chen et al. [11] and Saito et al. [12] to the FPN configuration for a fair comparison by applying the corresponding domain classifier to each FPN-level. Experiments using dedicated domain classifiers for each FPN-level showed no benefit over this strategy. For instance segmentation, we use the same loss functions for \mathcal{L}_{det} and \mathcal{L}_{att} as well as the same optimizer, learning rate, training schedule and batch size as in [1]. For monocular 3D detection, we again adopt the hyperparameters and loss functions for \mathcal{L}_{det} and \mathcal{L}_{att} from [36]. However, we use a reduced batch size of 8 due to memory constraints and a longer schedule, training for a total of 120k iterations and decaying the learning rate by a factor of 0.1 at 72k and 90k iterations.

B. Main Results

Instance Segmentation: For instance segmentation we consider the adaptation scenarios SYN \rightarrow CS, VIP \rightarrow CS and CS \rightarrow FCS. As shown in Tab. I, models trained in our setting outperform the unsupervised baselines by a large margin. This is also mirrored in the qualitative results shown in



Fig. 5: **Failure cases.** We show failure cases for 3D detection (SYN \rightarrow CS) and instance segmentation (VIP \rightarrow CS) in (a)-(c) and (d), respectively. (a): Due to the absence of one-ways in Synscapes, the orientations of all vehicles on the left hand side are flipped. (b): The lack of steep roads in the source domain results in a pitch bias. (c): Some vehicle types, e.g. vans, are underrepresented in the source domain, causing spurious predictions. (d): Contrary to Cityscapes, vehicle windows are not included in the segmentation masks in VIPER, leading to false negatives.

Fig. 3, where our approach yields more defined and accurate segmentations. While our approach is competitive with the fully supervised oracle for SYN \rightarrow CS and CS \rightarrow FCS, there remains a performance gap of 5.8 percentage points for VIP \rightarrow CS, despite substantial gains over the baselines. Here, the largest discrepancy (13.1 percentage points) is observed in the car class, which we attribute to an incompatibility in labeling policies: contrary to Cityscapes, vehicle segmentation masks in VIPER do not include windows, leading to false negatives in the target domain, which is illustrated in Fig. 5. Furthermore, we find that class-wise feature alignment does not improve the results for instance segmentation. This suggests that joint training of the base network does indeed promote a domain-invariant representation, which the instance segmentation module in turn reuses. Moreover, we observe that the unsupervised baseline methods only achieve a performance gain over the source only baseline in the CS \rightarrow FCS scenario, where the distribution of classes and scene layouts is identical. In both remaining scenarios they do not significantly improve the performance. We attribute this to negative transfer effects resulting from the class-agnostic alignment of features for domains with distinct distributions over scene layouts and relative frequencies of classes.

Monocular 3D detection: For monocular 3D detection we consider the adaptation scenarios SYN \rightarrow CS and CS \rightarrow FCS. From Tab. II we see that, compared to instance segmentation, monocular 3D detection is significantly more difficult: in this case, the simple joint training strategy still outperforms the source only and unsupervised baselines, but is unsuccessful in fully closing the gap towards the oracle. This suggests that for this task, the attribute network does not rely as strongly on the features learned via joint training. We observe that here, additionally performing feature alignment does improve performance, which supports this hypothesis. A possible explanation is that instance segmentation inherently synergizes more strongly with the detection task, as both estimate object representations in image space (i.e. 2D bounding boxes and instance masks), while monocular 3D detection estimates object representations in 3D space (i.e. 3D bounding boxes). Furthermore, class-wise feature alignment does indeed outperform its class-agnostic counterpart, verifying the benefit of utilizing the ground-truth class information available in our setting. As indicated by the qualitative results

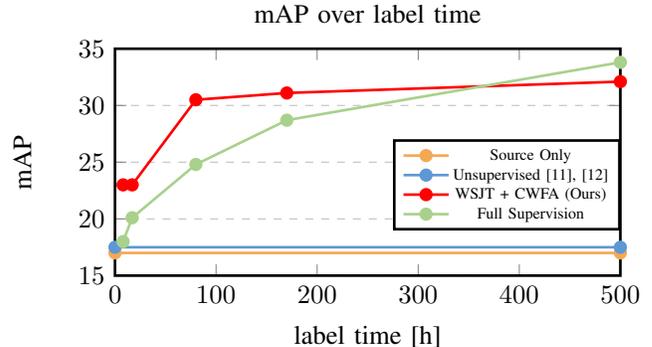


Fig. 6: **Ablation Study.** Instance segmentation performance (mAP) over label time (SYN \rightarrow CS).

in Fig. 4, our approach yields more accurate pose estimations compared to the baselines, but fails for scenes which have a layout that differs from those found in the source domain, e.g. one-way streets (Fig. 5).

C. Ablation Study

Performance over amount of weak target domain annotations: In Fig. 6, we compare the mAP over annotation time when using our setting compared to full supervision with annotated instance masks in the target domain. In internal experiments we observe that annotating the full task takes six times longer than annotating 2D bounding boxes for both 3D detection and instance segmentation, which for the latter matches previous reports [47]. The results show that up to a critical amount of labeled images, our approach is more efficient compared to fully supervised methods.

Ground-truth 2D boxes at test-time: One of the main questions in this work is if the network is able to learn transferable features for the cascaded detection module or if any observed performance gain is simply due to improved 2D detections. In Tab. III, we therefore analyze the performance of the instance segmentation and 3D box estimation modules when decoupled from the performance of the 2D detector by using ground-truth 2D bounding boxes as input during testing. While for instance segmentation joint training already yields good results, the 3D bounding box detector strongly benefits from the class-wise feature alignment. This supports

our hypothesis that for instance segmentation, joint training already learns transferable features, while for monocular 3D detection further improvements can be made by explicitly aligning the feature distributions.

V. CONCLUSION

In this work, we have presented a novel weakly-supervised domain adaptation setting, which exploits the structure of cascaded detection tasks. In our experiments, we have demonstrated that models adapted in our setting outperform unsupervised adaptation approaches by a large margin and can be competitive with their fully supervised counterparts. Although we considered instance segmentation and monocular 3D detection as examples, we are convinced this setting can be applied to other cascaded detection tasks such as human pose estimation, object tracking or trajectory forecasting.

REFERENCES

- [1] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *ICCV*, 2017.
- [2] R. Alp Gler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *CVPR*, 2018.
- [3] H. Zhao, R. Tachet des Combes, K. Zhang, and G. J. Gordon, "On learning invariant representations for domain adaptation," in *ICML*, 2019.
- [4] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *NIPS*, 2018.
- [5] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *CVPR*, 2019.
- [6] X. Jiang, Q. Lao, S. Matwin, and M. Havaei, "Implicit class-conditioned domain alignment for unsupervised domain adaptation," in *ICML*, 2020.
- [7] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015.
- [8] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *CVPR*, 2017.
- [9] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *ICML*, 2018.
- [10] B. Coors, A. P. Condurache, and A. Geiger, "Nova: Learning to see in novel viewpoints and domains," in *3DV*, 2019.
- [11] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," in *CVPR*, 2018.
- [12] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *CVPR*, 2019.
- [13] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-domain detection via graph-induced prototype alignment," in *CVPR*, 2020.
- [14] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, 2010.
- [15] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.
- [16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.
- [17] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fncs in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv:1612.02649*, 2016.
- [18] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv:1412.3474*, 2014.
- [19] B. Sun and K. Saenko, "Deep CORAL: correlation alignment for deep domain adaptation," *arXiv:1607.01719*, 2016.
- [20] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *ICCV*, 2017.
- [21] Q. Lian, F. Lv, L. Duan, and B. Gong, "Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach," in *ICCV*, 2019.
- [22] Z. He and L. Zhang, "Multi-adversarial faster-rcnn for unrestricted object detection," in *ICCV*, 2019.
- [23] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *CVPR*, 2020.
- [24] Y. Zheng, D. Huang, S. Liu, and Y. Wang, "Cross-domain object detection through coarse-to-fine feature adaptation," in *CVPR*, 2020.
- [25] C. Hsu, Y. Tsai, Y. Lin, and M. Yang, "Every pixel matters: Center-aware feature alignment for domain adaptive object detector," in *ECCV*, 2020.
- [26] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *CVPR*, 2019.
- [27] T. Wang, X. Zhang, L. Yuan, and J. Feng, "Few-shot adaptive faster r-cnn," in *CVPR*, 2019.
- [28] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready, "A robust learning approach to domain adaptive object detection," in *ICCV*, 2019.
- [29] S. Kim, J. Choi, T. Kim, and C. Kim, "Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection," in *ICCV*, 2019.
- [30] P. Zama Ramirez, A. Tonioni, S. Salti, and L. Di Stefano, "Learning across tasks and domains," in *ICCV*, 2019.
- [31] B. Zhou, N. Kalra, and P. Krähenbühl, "Domain adaptation through task distillation," in *ECCV*, 2020.
- [32] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," in *TIP*, vol. 28, 2019.
- [33] T. Gebru, J. Hoffman, and L. Fei-Fei, "Fine-grained recognition in the wild: A multi-task domain adaptation approach," *ICCV*, 2017.
- [34] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *CVPR*, 2018.
- [35] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3d human pose estimation in the wild: A weakly-supervised approach," in *ICCV*, 2017.
- [36] A. Simonelli, S. R. Buló, L. Porzi, M. López-Antequera, and P. Kotschieder, "Disentangling monocular 3d object detection," in *ICCV*, 2019.
- [37] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017.
- [38] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.
- [39] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *ECCV*, 2018.
- [40] M. Wrenninge and J. Unger, "Synscapes: A photorealistic synthetic dataset for street scene parsing," *arXiv:1810.08705*, 2018.
- [41] S. R. Richter, Z. Hayder, and V. Koltun, "Playing for benchmarks," in *ICCV*, 2017.
- [42] N. Ghlert, N. Jourdan, M. Cordts, U. Franke, and J. Denzler, "Cityscapes 3d: Dataset and benchmark for 9 dof vehicle detection," *arXiv:2006.07864*, 2020.
- [43] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015.
- [44] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.
- [45] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *arXiv:1903.11027*, 2019.
- [46] L. Qi, L. Jiang, S. Liu, X. Shen, and J. Jia, "Amodal instance segmentation with kins dataset," in *CVPR*, 2019.
- [47] M. Bellver, A. Salvador, J. Torres, and X. Giró-i-Nieto, "Budget-aware semi-supervised semantic and instance segmentation," in *CVPR Workshops*, 2019.