Cross-Domain Object Detection for Autonomous Driving: A Stepwise Domain Adaptative YOLO Approach

Guofa Li, *Member, IEEE*, Zefeng Ji, Xingda Qu, Rui Zhou, and Dongpu Cao

***Abstract*—Supervised object detection models based on deep learning technologies cannot perform well in domain shift scenarios where annotated data for training is always insufficient. To this end, domain adaptation technologies for knowledge transfer have emerged to handle the domain shift problems. A stepwise domain adaptive YOLO (S-DAYOLO) framework is developed which constructs an auxiliary domain to bridge the domain gap and uses a new domain adaptive YOLO (DAYOLO) in cross-domain object detection tasks. Different from the previous solutions, the auxiliary domain is composed of original source images and synthetic images that are translated from source images to the similar ones in the target domain. DAYOLO based on YOLOv5s is designed with a category-consistent regularization module and adaptation modules for image-level and instance-level features to generate domain invariant representations. Our proposed method is trained and evaluated by using five public driving datasets including Cityscapes, Foggy Cityscapes, BDD100K, KITTI, and KAIST. Experiment results demonstrate that object detection performance is significantly improved when using our proposed method in various domain shift scenarios for autonomous driving applications.**

***Index Terms*—Autonomous vehicles, adversarial learning, deep learning, domain adaptation, object detection.**

I. Introduction

OBJECT detection is an essential task for advanced driver assistance systems (ADASs) and autonomous vehicles (AVs) aiming to classify and localize objects of interest in images. The recent development of deep learning technologies has enabled the emergence of several state-of-the-art (SOTA) models for object detection, which has greatly improved object detection performance [1]–[4]. However, approaches based on deep learning are mostly data-driven, which means that these approaches usually rely on large-scale annotated datasets for supervised learning [5], [6]. This limits the ability of the deep models to generalize to new surroundings (i.e., the target domain) where the image style, resolution, and even illumination or weather conditions diverge significantly from the training images (i.e., the source domain). Although the training of deep models with additional data collected from the target domain can effectively improve their performance, the huge overhead of labeling work makes it not always feasible to collect large-scale annotated training data from new surroundings.

It has been reported that the domain shift problems (i.e., data bias between training data and test data) can cause significant performance degradation of deep models [7]. Unsupervised domain adaptation provides an attractive solution to addressing domain shift problems by adapting the deep models from label-rich source samples to label-poor target samples. In recent years, researchers have extensively applied domain adaptation methods to resolve image classification problems [8], [9]. However, the related knowledge in object detection tasks is still lacking and needs to be enhanced.

In order to handle domain shift problems, a novel stepwise domain adaptive YOLO framework (S-DAYOLO) for cross-domain object detection is proposed. An unpaired image-to-image translator is firstly trained to construct synthetic images. Domains containing the synthetic images and the source domain images are defined as the auxiliary domain, which aim to simplify the process of aligning different domains. The proposed domain adaptive YOLO (DAYOLO) based on YOLOv5s [10] integrates a category-consistent regularization module and adaptation modules for image-level and instance-level features to reduce the $\mathcal{H}$-distance between domains. Specifically, in the adaptation modules for image-level and instance-level features, domain classifiers are trained with adversarial strategy to generate domain-invariant representations. The category-consistent regularization module is further introduced to activate the category-specific features of the object-related area and to mine the hard aligned category. This category-consistent regularization module is novel in the literature and has never been reported before.

In summary, the main contributions of this study are as follows: 1) A novel auxiliary domain based on source and synthetic samples is proposed to bridge the domain gap, which improves the ability of detectors to generate domain-invariant

Manuscript received 18 February 2022; revised 23 March 2022; accepted 4 April 2022. Date of publication 6 April 2022; date of current version 24 October 2022. This work was supported by Shenzhen Fundamental Research Fund under Grant JCYJ20190808142613246. *(Corresponding author: Xingda Qu.)*

Guofa Li, Zefeng Ji, and Xingda Qu are with the Institute of Human Factors and Ergonomics, College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China (e-mail: hanshan198@gmail.com; jizefeng0810@163.com; quxd@szu.edu.cn).

Rui Zhou is with Waytous Inc., Beijing 999077, China (e-mail: rui.zhou@qaii.ac.cn).

Dongpu Cao is with the School of Vehicle and Mobility, Tsinghua University, 100084 Beijing, China (e-mail: dp_cao2016@163.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2022.3165353.

Digital Object Identifier 10.1109/TIV.2022.3165353

representations for better detection performance in both source and target domains; 2) A novel DAYOLO model is designed with a category-consistent regularization module and adaptation modules for image-level and instance-level features, which aligns feature distributions between the auxiliary and target domains to learn an advanced cross-domain object detector; 3) Extensive ablation experiments are conducted on domain shift scenarios to evaluate the cross-domain detection performance of S-DAYOLO. This allows us to determine the generalizability of our method in all-around-the-clock illuminations in various weather conditions.

II. Related Work

A. Object Detection

The success of deep convolutional neural networks (DCNN) has greatly promoted the rapid shift paradigm in object detection over the past few years [11]. The frameworks of mainstream anchor-based object detector are divided into two types [12]: two-stage object detector (e.g., Faster RCNN [13]) and one-stage object detector (e.g., YOLO[14] and SSD [15]). Among the large number of models for object detection, the one-stage object detector YOLO has received extensive attention due to its effectiveness and real-time performance. In this study, YOLOv5s is chosen as the base detection model, and improve its detection performance in cross-domain scenarios.

B. Deep Domain Adaptation

Domain adaptation is utilized to tackle domain shift problems between a label-rich source domain and a label-poor target domain by reducing the feature distribution gap between domains. DCNN can generate more transferable representations by unlocking explanation factor of inter-domain variations [16]. Deep domain adaptive networks for image classification which have received extensively attention in recent years usually extracts domain invariant representations by embedding domain adaptative modules in deep networks [8], [9]. Recent studies have shown that the popular approaches for domain adaptation include statistical moment matching based approaches (i.e., maximum mean discrepancy [17], central moment discrepancy [18], and second-order statistics matching [19]) and adversarial loss based approaches [20]. Several recent studies [21] have also proposed pixel-level domain adaptation for unpaired image translation between two domains. Unlike these studies, we focus on the more challenging object detection tasks where the objects' locations and categories are unknown and need to be predicted.

C. Domain Adaptive Object Detection (DAOD)

The objective of DAOD is to learn strongly generalize models for cross-domain object detection using label-rich source samples and label-poor target samples. The training process relies on models or principles of domain adaptation. Chen *et al.* [7] proposed a Domain Adaptive Faster RCNN (DAF) framework and are about the first to formulate and address the problem of DAOD. Chen *et al.* used $\mathcal{H}$ distance to measure the divergence between the two domains and aligned feature distributions in an adversarial training manner. A Selective Cross-domain Alignment (SCDA) framework proposed by Zhu *et al.* [22] selectively align the features between domains by mining discriminative regions. He *et al.* [23] proposed Multi-Adversarial FasterRCNN (MAF) that employs a multi-adversarial domain classifier sub-module to address unrestricted object detection problems. He *et al.* narrowed the feature map without losing information to improve the training efficiency of the detector. Collaborative Self-training (CST) was developed by Zhao *et al.* [24] to train the region proposal network and region proposal classifier. The customized maximizing discrepancy proposed by Zhao *et al.* effectively exploits low-confidence regions of interest to further improve the precision and generalizability of model. A conditional domain normalization (CDN) module was proposed by Su *et al.* [25] to map information from source and target samples into the public space for object detection. Zhang *et al.* [26] proposed a virtual-real interaction method to reduce the adverse effects of domain shift, which guides the model to extract common information between virtual and real data. Unlike previous studies, in this work, we firstly construct an auxiliary domain to bridge the domain gap, and then develop a DAYOLO model with an adversarial learning strategy to generate domain-invariant representations.

One of the relevant methods is a Multiscale Domain Adaptive YOLO (MS-DAYOLO) [27] framework proposed by Hnewa and Radha based on YOLOv4 [28], which uses multiple domain-adaptive paths and the corresponding domain classifiers at different scales of the backbone network to generate domain-invariant representations. MS-DAYOLO is different from our work mainly in two aspects. First, unlike MS-DAYOLO, our proposed model is developed based on DAYOLO which uses YOLOv5s to further align features at the instance level and design a category-consistent regularization module to reduce background activation for the mining of hard-to-align instance samples. Second, our work constructs an auxiliary domain to bridge the domain gap, which is different from MS-DAYOLO.

Some other techniques are also helpful to reduce the dependence of deep models on labeled data in new scenarios. Li *et al.* [29] collected synthetic images by simulating various realistic driving scenarios in Unity3D, and combined other public datasets to improve the accuracies of object detection algorithms. A parallel execution system with virtual and real scenes for scene-specific pedestrian detection was proposed by Zhang *et al.* [30], which generates data in a virtual scene and adapts the generic model to specific scenes.

III. Methodology

A. Distribution Alignment With $\mathcal{H}$ Distance

The auxiliary domain is denoted as $D_A = \{(x_i^A, y_i^A)\}_{i=1}^{N_A}$ and it consists of N_A number of images, where x_i^A denotes the i-th image and $y_i^A = (b_i^A, c_i^A)$ denotes the corresponding bounding box annotations b_i^A with category label c_i^A. The target domain $D_T = \{(x_i^T)\}_{i=1}^{N_T}$ has N_T number of images with no ground-truth annotations. In domain adaptation, the feature distributions between the auxiliary and target samples are assumed to be similar but different (i.e., $D_A \neq D_T$). The objective of

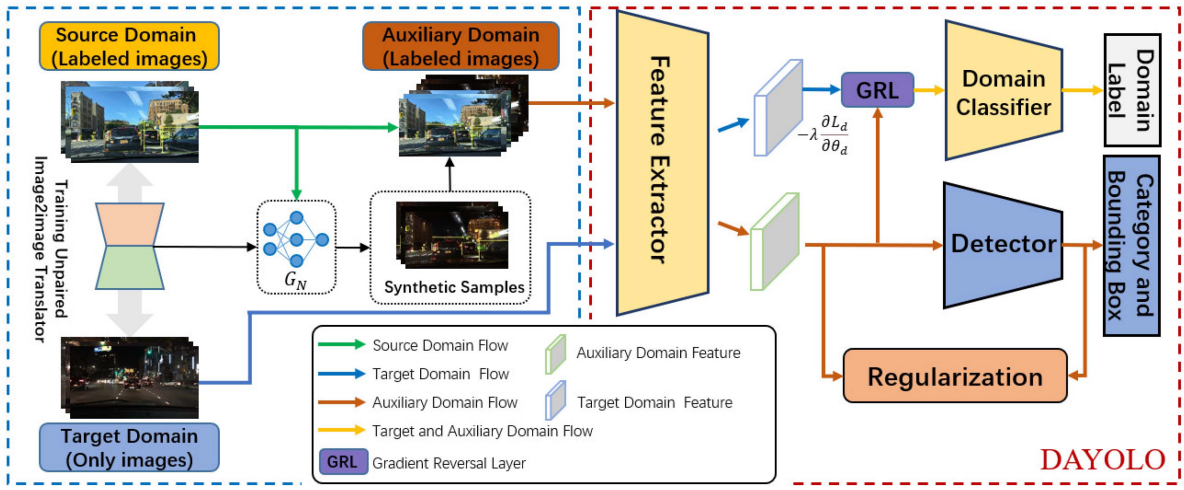


Fig. 1. The structure of S-DAYOLO for domain adaptation with both auxiliary sample generation and invariant feature learning. As illustrated in the blue dash box, auxiliary domain images consist of source domain images and synthetic images which are generated by a pretrained unpaired image-to-image translator model (i.e., CycleGAN [33]). Domain invariant feature learning of auxiliary and target domains is conducted with GRL and regularization modules as illustrated in the red dash box.

domain adaptation is to train models with good generalizability in the target domain using the data form the auxiliary and target domains.

To achieve this objective, a $\mathcal{H}$-distance is designed by Ben *et al.* [31] to measure the domain divergence with different data distributions. The auxiliary and target images are represented as $x_A \in A$ and $x_T \in T$, respectively. A domain classifier represented as $h(x) : x \to \{0, 1\}$ receives the input image $x \in \{A \cup T\}$ and predicts the domain of the input image. Considering $\mathcal{H}$ to be a group of domain classifiers, $\mathcal{H}$-distance can be defined as:

$$d_{\mathcal{H}}\ (A, T) = \ 2\left(1 - \min_{h\in\mathcal{H}} \left(\epsilon_A\left(h\left(x\right)\right) + \epsilon_T\left(h\left(x\right)\right)\right)\right), \quad (1)$$

where ϵ_A and ϵ_T denote the expected overall prediction errors for auxiliary and target domain samples, respectively. (1) shows that the larger the error rate of domain classifiers, the smaller $\mathcal{H}$-distance, and vice versa.

In DCNN, the network that produces x is denoted as f. To achieve the smaller $\mathcal{H}$-distance from auxiliary and target samples, the network f is enforced to output feature vectors to align the feature distribution of two domains, which results in:

$$\arg\ \min_f d_{\mathcal{H}}\left(A, T\right) \Leftrightarrow \ \max_f \min_{h\in\mathcal{H}} \left(\epsilon_A\left(h\left(f\left(x\right)\right)\right)\right. \\ \left. + \ \epsilon_A\left(h\left(f\left(x\right)\right)\right)\right), \quad (2)$$

The gradient reverse layer (GRL) that realizes the optimization of Eq. (2) in the adversarial training manner is proposed by Ganin *et al.* [32] and is developed with DCNN for image classification in cross-domain scenarios.

B. Framework Overview

The overall framework of S-DAYOLO for a universal object detector in domain shift scenarios is presented in Fig. 1. The whole framework composes of two stages: auxiliary domain construction and domain-invariant representations learning. In the auxiliary domain construction stage, CycleGAN [33] is employed to train an image-to-image translator by using the source and target images. Synthetic images are translated from source domain images so that they have the same content but different domains. The auxiliary domain data are composed of source domain images and synthetic images as illustrated by the blue dashed box in Fig. 1. The process of domain-invariant representations learning is illustrated by the red dashed box of Fig. 1. The domain classifiers are trained with adversarial strategy to generate domain-invariant representations. A regularization component is further introduced to activate the object-related area and mine the hard aligned category. In the training stage of the detector, several adaptative components are added so that the feature extractor learns domain-invariant representations for domain adaptation, while the inference phase is the same as the vanilla YOLOv5s.

C. Auxiliary Domain Construction

As mentioned above, CycleGAN [33] is utilized to train an unpaired image-to-image translator, which builds mapping functions between source and target samples by learning invariant content information and variant style information. Fig. 2 shows that a well-trained translator can perform conversions in different domains without changing the content information. CycleGAN that is based on adversarial learning forms a ring network through two paired generators and two independent discriminators. The total objective function is shown as:

$$L_{CycleGAN} = L_{GAN}\ \left(G_S^T, D_T, S, T\right) \\ + L_{GAN}\left(G_T^S, D_S, S, T\right) + \lambda L_{cyc}\left(G_S^T,\ G_T^S\right), \quad (3)$$

where S and T present the source and target domains, respectively, G_X^Y defines a mapping function that transform X domain to Y domain, and D_S and D_T denote discriminators in the

Fig. 2. Examples of synthetic samples generated by GycleGAN in different source and target domains: (a) and (b) Represent Cityscapes [34] image translated to Foggy Cityscapes [35] and BDD [36] daytime, respectively; (c) Shows KITTI [37] image translated to Cityscapes; (d) Shows BDD daytime image translated to BDD nighttime; (e) Shows KAIST [38] RGB translated to KAIST LWIR.

two different domains, respectively. L_{GAN} and L_{cyc} denote adversarial loss and cycle consistency loss, respectively. λ is a weight parameter for L_{GAN} and L_{cyc}. In our work, we only use the trained translator that converts source images to target images to generate synthetic images. As the translator performs the conversion between different domains without changing the content information, the synthetic and source images share the same annotations. To reduce the impact of imperfect image-to-image translation on object detection models and improve the generalizability of object detection models between the source and target domains, synthetic and source samples jointly construct auxiliary domains for domain-invariant representations learning.

D. Domain-Invariant Representations Learning

The detail of our proposed DAYOLO structure for domain invariant representations learning is presented in Fig. 3. DAYOLO that learns domain-invariant representations between the auxiliary and target domains can be divided into a category-consistent regularization module and adaptation modules for image-level and instance-level features:

1) Image-Level Adaptation Module: The image-level feature refers to the feature vector with shallow semantics such as image background and scene layout in the basic detection network. As shown by the structure of DAYOLO in Fig. 3, the three different scale feature vectors of the basic detection network are fed to the image-level adaptation module. It consists of two convolutional layers to discriminate the domain category (i.e., auxiliary or target domain), and a focal loss with a penalty is used to calculate the images-level adaptation loss L_{img}, as follows:

$$L_{img} = -\frac{1}{N}\sum_{ixy}\begin{cases}\alpha(1-\hat{Y}_{ixy})^{\gamma}\log\hat{Y}_{ixy}, & if \quad Y_{ixy}=1\\ (1-\alpha)\,(\hat{Y}_{ixy})^{\gamma}\log\left(1-\hat{Y}_{ixy}\right), & otherwise\end{cases} \tag{4}$$

where γ and α are assigned to 1.5 and 0.25, respectively, N is the number of samples in training stage, Y_{ixy} denotes the domain category of the i-th input image at location (x, y) of the feature map with auxiliary domain as label 1 and target domain as label 0, and $\hat{Y}_{ixy}$ denotes the domain category probability output for the i-th input image at location (x, y) of the feature map.

As discussed in Section III. A, to reduce the $\mathcal{H}$ distance, GRL is implemented to solve the problem that joint maximization and minimization. To achieve the smaller $\mathcal{H}$-distance between the auxiliary and target domains, the parameters of the basic detection network need to be optimized to maximize the image-level adaptation loss L_{img}. Differently, to learn domain classifiers, the parameters of domain classifiers need to be optimized to minimize L_{img}.

2) Instance-Level Adaptation Module: The instance-level feature refers to the feature vector of the area related to the instance before the detection head. As shown by the structure of DAYOLO in Fig. 3, the instance-level adaptation module is implemented before the detection head to reduce local instance divergence (e.g., instance appearance, size, and viewpoint). Similarly, domain classifiers with GRL are used to reduce the $\mathcal{H}$-distance to align instance-level features. The instance-level adaptation loss L_{ins} is computed as follows:

$$L_{ins} = -\sum_{i}\left[p_i\log\hat{p}_i + (1-p_i)\log\left(1-\hat{p}_i\right)\right], \tag{5}$$

where p_i and $\hat{p}_i$ denote the domain category label and the probability output of classifiers for the i-th image, respectively.

3) Category-Consistent Regularization Module: To activate the category-specific features of the object-related area, as illustrated in Fig. 3, a multi-label image classifier is attached to the back of the spatial pyramid pooling (SPP) layer and trained with supervisions from the auxiliary domain. The multi-label image classifier is designed as two convolutional layers with filters of size 3×3 and a fully connection layer. The numbers of filters are 256 and 64 in the two convolutional layers, respectively. The multi-label image classifier loss L_{ML} is computed as follows:

$$L_{ML} = -\sum_{c=1}^{C} y^c\log\left(\hat{y}^c\right) + \left(1-y^c\right)\log\left(1-\hat{y}^c\right), \tag{6}$$

where C represents the number of categories, $\hat{y}^c$ represents the probability output of the c-th category, and $y^c \in \{0, 1\}$ represents the label of the c-th category.

The category-consistent regularization module is used to mine the hard aligned category in domain adaptation. As the multi-label image classifier explores the whole image-level context, and the detection head has more accurate instance-level object area features, the former and latter are complemental. Therefore, the consistency between the output of the multi-label image classifier and the detection head is utilized as a measure for the difficulty of a certain category, and also for the weight of the hard-aligned category during instance-level adaptation. Specifically, the distance function that measures the categorial

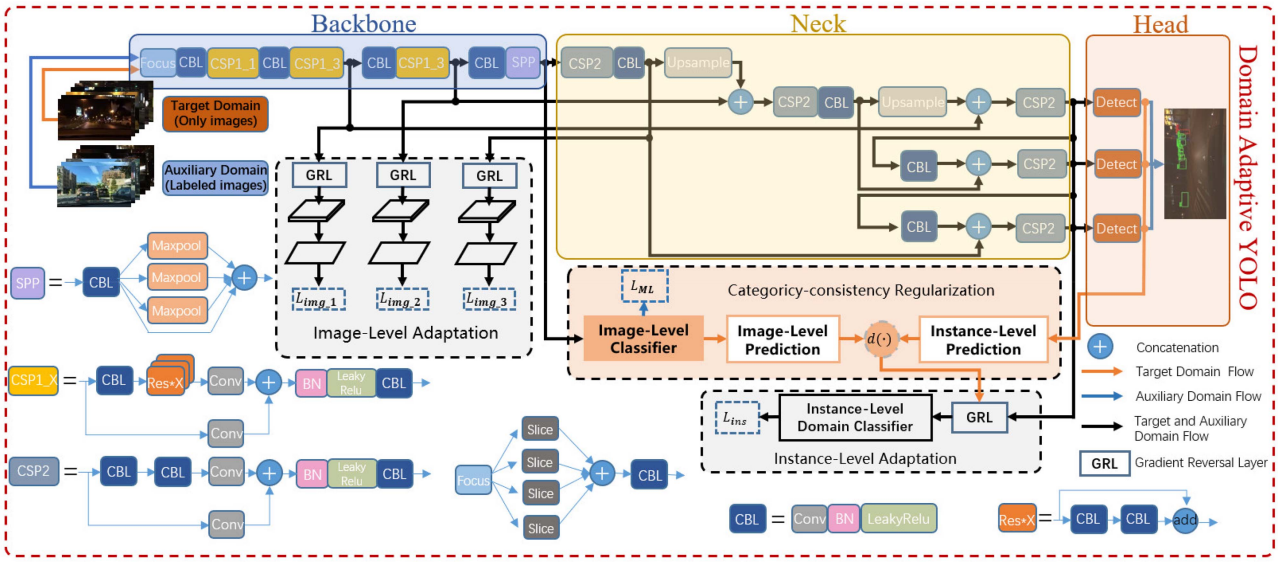


Fig. 3. The detailed architecture of DAYOLO.

consistency between the output of the multi-label image classifier and the detection head is defined as,

$$d_i = \frac{1}{C} \sum_{c=1}^{C} |\hat{p}_i^c - \hat{y}_i^c| , \tag{7}$$

where $\hat{y}_i^c$ is the probability output of the c-th category for the i-th training image from the multi-label image classifier, $\hat{p}_i^c$ denotes the maximum estimated probability output of the c-th category for the i-th training image from the detection head. (7) is utilized to weigh the instance-level adaptation loss. The instance-level adaptation loss with category-consistent regularization can be written as:

$$L_{ins}^{Reg} = -\sum_i d_i \left[p_i \log \hat{p}_i + (1 - p_i) \log (1 - \hat{p}_i) \right] , \tag{8}$$

where d_i is only applied as the weights of instance-level adaptation loss for target samples. As auxiliary samples has supervision signals (i.e., annotated labels), the weights for the source domain keep unchanged (i.e., $d_i = 1$).

E. Objective Function

The objective function is made up of detection loss L_{det}, image-level adaptation loss L_{img}, instance-level adaptation loss with category-consistent regularization L_{ins}^{Reg}, and multi-label loss L_{ML}. The network can automatically classify and locate objects of interest from images by optimizing L_{det} and can activate the category-specific features of the object-related area by optimizing L_{ML}. The goal of L_{img} is to achieve the smaller $\mathcal{H}$-distance of image-level features between auxiliary and target domains. The goal of L_{ins}^{Reg} is to mine the hard aligned category and align the instance-level feature distributions between auxiliary and target domains. To compute the detection loss L_{det} and multi-label loss L_{ML}, only auxiliary domain images with the ground-truth annotations are used. The total objective function of DAYOLO is expressed as:

$$L = L_{det} + L_{ML} + \lambda \left(L_{img} + L_{ins}^{Reg} \right) , \tag{9}$$

where λ is a trade-off parameter and defaults to 0.1. No additional hyper parameters are introduced to DAYOLO. For the implementation in domain adaptation components, the GRL which automatically reverses the gradient during back propagation is used to implement adversarial learning to generate domain-invariant representations. In the inference process, the domain adaptive components are removed and the vanilla YOLOv5s architecture with adaptive weights is used, which implies that our method will not slow down the inference speed when the model is deployed.

IV. Experiments and Results

A. Datasets and Scenarios

Datasets: In experiments, public driving datasets including Cityscapes [34], Foggy Cityscapes [35], BDD100K [36], KITTI [37], and KAIST [38] are used to set domain shift scenarios.

1) *Cityscapes:* Object detection under complex urban street scenarios is an important issue with broad application prospects for AVs. The Cityscapes dataset is mainly captured in clear weather, and it provides annotations for instance segmentation. Its training set and validation set have 2975 and 500 images, respectively.
2) *Foggy Cityscapes:* Object detection on urban streets in adverse weather is a challenging problem for AVs. The foggy Cityscapes dataset are created by simulating fog over the Cityscapes dataset. Thus, the number of images in Foggy Cityscapes is the same as Cityscapes. The foggy images from Foggy Cityscapes with a visibility of 150 meters are used in experiments.
3) *KITTI:* The KITTI dataset which has 7481 training images is constructed for the development of AVs and mobile

TABLE I
NUMBER OF CATEGORY INSTANCES IN THE REFINED BDD100K DATASET

Dataset		Person	Rider	Car	Bus	Truck	Bicycle	Motor	Train
Training	Day	14,416	951	74,987	1,867	4,392	1,321	692	24
	Night	14,652	651	146,138	1,368	3,002	1,177	487	17
	Rainy	3,829	110	16,286	494	1,158	231	53	8
Validation	Day	2,146	124	10,681	259	610	163	73	0
	Night	2,124	86	20,908	196	403	125	66	1
	Rainy	446	18	2,422	69	284	22	15	0

robotics. The KITTI dataset contains hours of videos from traffic scenes, which are recorded by various high-quality sensors (i.e., RGB, grayscale, and depth sensors). The dataset comprises of 28742 cars with annotations.

4) *BDD100K:* The Berkeley Deep Drive (BDD) dataset with 100000 images is collected in a large diversity environment in terms of scenes and geographical regions [39]. In our experiment, the images of clear weather and city street scenes from BDD100K are filtered and retained. With further refinement applied, images with attribute of daytime and nighttime are selected. Besides, rainy images collected during daytime in city street scenes are used to examine the effectiveness of our proposed method in Sunny → Rainy (i.e., Cityscapes → BDD rainy) adaptation. The detailed numbers of the category instances in the refined BDD100K dataset are presented in Table I.
5) *KAIST:* The KAIST dataset which contains RGB and long-wavelength infrared (LWIR) images is a dataset collected around the clock for pedestrian detection and multi-source fusion [40]. As images between consecutive frames are highly similar, the dataset is refined by following the method proposed in [41]. Moreover, the images with no pedestrians are removed. The refined KAIST dataset has 6294 images for training and 579 images for testing, with 11058 and 1161 labeled pedestrians, respectively.

Scenarios: Four domain shift driving scenarios for cross-domain object detection are set to evaluate our proposed S-DAYOLO: 1) adverse weather adaptation (Cityscapes → Foggy Cityscapes and Cityscapes → BDD rainy), where source images are captured in clear weather, while target images are captured in adverse weather; 2) cross-camera adaptation (KITTI → Cityscapes and Cityscapes → BDD daytime), where the source and target domain data are taken with different camera setups; 3) daytime-to-nighttime adaptation (BDD daytime → BDD nighttime), where the source and target samples are captured in daytime and nighttime, respectively; 4) heterogeneous adaptation (KAIST RGB → KAIST LWIR), where source samples are RGB images, while target samples are long-wavelength infrared images. The symbol '→' indicates the direction of adaptation from the source domain to the target domain. The information of the dataset is given in Table II.

TABLE II
THE NUMBER OF IMAGES IN DIFFERENT DOMAIN SHIFT SCENARIOS. (C, F, KI, D, N, R, KR AND KL REPRESENT CITYSCAPES, FOGGY CITYSCAPES, KITTI, BDD DAYTIME, BDD NIGHTTIME, BDD RAINY, KAIST RGB, AND KAIST LWIR, RESPECTIVELY. THE SYMBOL '→' MEANS THE DIRECTION OF ADAPTATION.)

Scenarios	Training set		Validation set
	Source Domain	Target Domain	Target Domain
$C \rightarrow F$	Cityscapes	Foggy Cityscapes	Foggy Cityscapes
	2,975	2,975	500
$C \rightarrow R$	Cityscapes	BDD rainy	BDD rainy
	2,975	3,280	241
$KI \rightarrow C$	KITTI	Cityscapes	Cityscapes
	7,481	2,975	500
$C \rightarrow D$	Cityscapes	BDD daytime	BDD daytime
	2,975	6,647	933
$D \rightarrow N$	BDD daytime	BDD nighttime	BDD nighttime
	6,647	15,090	2,133
$KR \rightarrow KL$	KAIST RGB	KAIST LWIR	KAIST LWIR
	6,294	6,294	579

B. Experimental Setup

The auxiliary domain is constructed by training CycleGAN with data from the source and target domains. DAYOLO is initialized with the weights trained in the ImageNet dataset. Following the default settings in [10], a Stochastic Gradient Descent optimizer (SGD) is utilized to optimize the parameters of model. The learning rate, weight decay and momentum of the SGD optimizer are assigned to 0.01, 0.0005 and 0.937, respectively. The data fed to the network consists of two sets: data with images and annotations (i.e., bounding boxes and category) from auxiliary domain, and data without annotations from target domain. The number of each batch is assigned to 64 during training. Each batch is composed of 32 auxiliary and target samples. The experiments are carried out with Pytorch and are implemented in an Intel-i9 10920X CPU (3.50 GHz) with 64GB of RAM, and 2 NVIDIA RTX3090 GPUs with 24 GB of memory. For experimental results, average precision (AP) results and mean average precision (mAP) results with an IoU threshold of 0.5 are presented by using the validation sets with annotations from the target domains.

C. Experimental Results

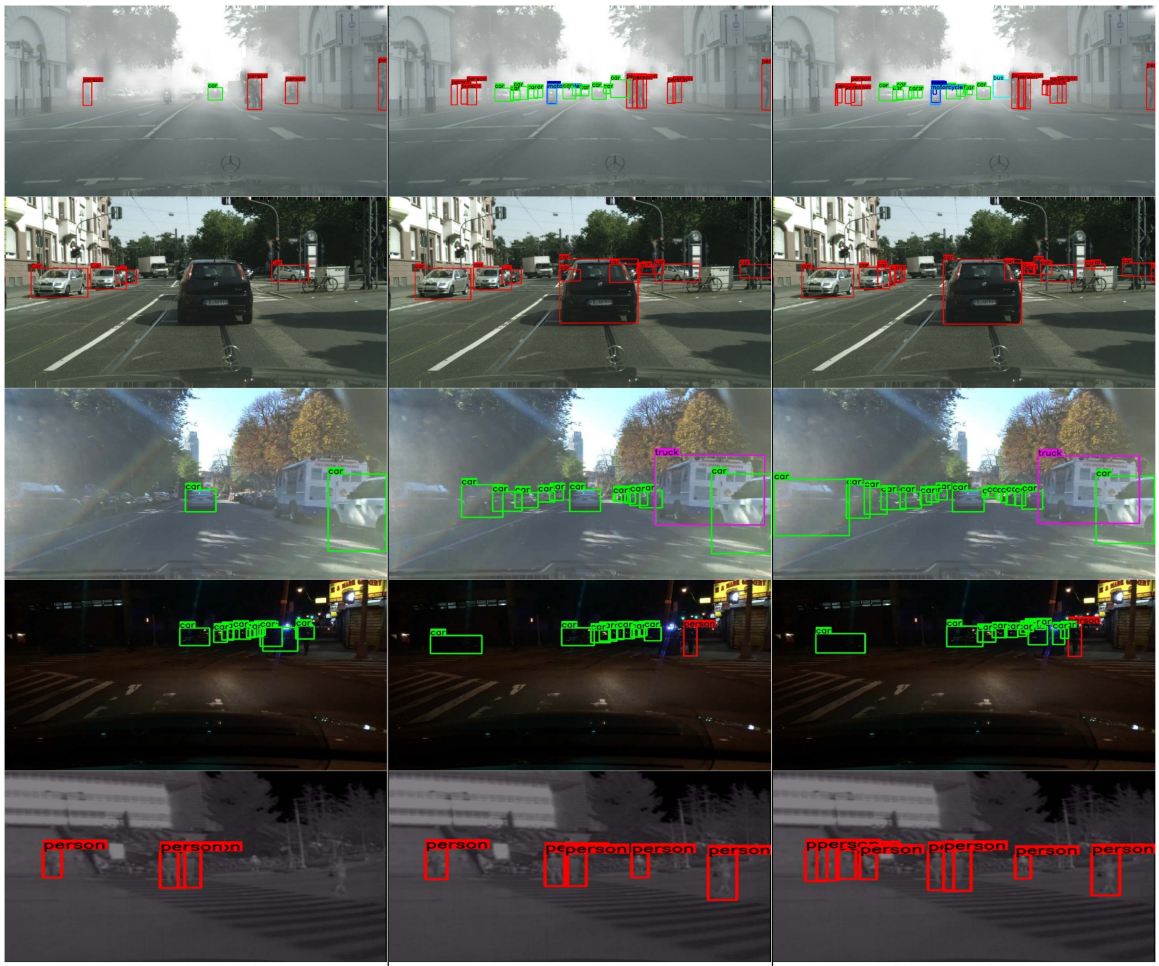
Our proposed S-DAYOLO is evaluated on different domain shift scenarios for object detection by comparing with six SOTA domain adaptation methods including DAF [7], SCDA [22], MAF [23], CST [24], CDN [25] and MS-DAYOLO[27]. The results are shown as follows.

a) Adverse Weather Adaptation: Since adverse weather cannot be avoided in real world traffic, stable object detection performance in adverse weather is important for safety-critical applications such as ADASs and AVs. In this subsection, S-DAYOLO is evaluated by adapting clear weather (i.e., Cityscapes) to foggy weather (i.e., Foggy Cityscapes).

Table III summarizes the quantitative results of the examined methods, which demonstrates that S-DAYOLO improves the detection performance on the validation set of Foggy Cityscapes, and achieves the best performance. Specifically, the mAP result of S-DAYOLO is improved by at least 2.4% by comparing with the other five SOTA methods. Compared with the Source-Only model, S-DAYOLO improves the mAP by 17.5%, which means that our method can significantly alleviate the domain shift problem between the examined weathers. Ablation studies are

TABLE III
THE RESULTS OF MAP (%) FOR ADVERSE WEATHER ADAPTATION TASK. THE METHODS ARE TRAINED USING SOURCE SAMPLES FROM CITYSCAPES AND TARGET SAMPLES FROM FOGGY CITYSCAPES. "ORACLE" DENOTES A YOLOV5S MODEL WITH THE SAME SETTING SUPERVISED TRAINED WITH TARGET SAMPLES

Method	Person	Rider	Car	Truck	Bus	Motor	Bike	Train	mAP
Source-Only	30.2	31.2	38.1	10.2	23.4	11.1	25.4	10.7	22.5
DAF [7]	25.0	31.0	40.5	22.1	35.3	20.0	27.1	20.2	27.5
SCDA [22]	33.5	38.0	48.5	26.5	39.0	28.0	33.6	23.3	33.8
MAF [23]	28.2	39.5	43.9	23.8	39.9	29.2	33.9	33.3	34.0
CST [24]	32.7	44.4	50.1	21.7	**45.6**	30.1	36.8	25.4	35.9
CDN [25]	35.8	**45.7**	50.9	**30.1**	42.5	**30.8**	36.5	29.8	36.6
Ours w/o DAYOLO	40.3	41.7	59.5	21.8	37.0	30.6	35.7	38.6	38.1
Ours (DAYOLO)	38.5	37.1	52.2	17.6	31.2	15.1	34.3	26.0	31.5
Ours (S-DAYOLO)	**42.6**	42.1	**61.9**	23.5	40.5	24.4	**37.3**	**39.5**	**39.0**
Oracle	44.9	45.2	66.7	32.7	52.7	35.3	39.4	51.8	46.1



(a) Source-Only (b) Ours (c) Ground Truth

Fig. 4. Examples of the qualitative detection results. (a), (b), and (c) Display the results on the target domain using Source-Only model, S-DAYOLO model and ground truth labels, respectively. The first row illustrates the detection results in Cityscapes $\rightarrow$ Foggy Cityscapes. The second and third rows present the detection results on KITTI $\rightarrow$ Cityscapes and Cityscapes $\rightarrow$ BDD daytime, respectively. The fourth row presents the detection results in daytime-to-nighttime adaptation. The last row illustrates the heterogeneous adaptation detection results.

conducted for further investigation. When only the auxiliary domain is introduced, denoted as "Ours w/o DAYOLO", the mAP result improves by 15.6% compared to the Source-Only model. Compared with "Ours w/o DAYOLO", DAYOLO can further assist the adaptive process by generating domain-invariant representations and improve the result of mAP by 0.9%. The first row in Fig. 4 shows examples of detection results when using S-DAYOLO on the Foggy Cityscapes dataset as compared to the Source-Only model and the ground truth labels.

b) Cross-Camera Adaptation: Since the images captured by different cameras differ in quality, scales, and viewing angles, data bias inevitably exists in data collection which can lead to

TABLE IV
AVERAGE PRECISION (AP) RESULTS (%) OF CAR ON THE CITYSCAPES VALIDATION SET IN KITTI $\rightarrow$ Cityscapes. "ORACLE" DENOTES A YOLOV5S MODEL WITH THE SAME SETTING SUPERVISED TRAINED WITH TARGET SAMPLES

Method	Car AP
Source-Only	23.7
DAF [7]	38.5
SCDA [22]	42.5
MAF [23]	41.0
CST [24]	43.6
CDN [25]	44.9
Ours w/o DAYOLO	36.5
Ours (DAYOLO)	48.7
Ours (S-DAYOLO)	**49.3**
Oracle	72.1

domain shift problems. Two domain shift scenarios are used to evaluate our proposed method in this regard, namely $KITTI \rightarrow Cityscapes$ and $Cityscapes \rightarrow BDD\ daytime$.

The quantitative results of detection performance are presented in Tables IV and V. From the results shown in Table IV, there is a significant performance gap between the Source-Only and Oracle models. Compared with the Source-Only model, our proposed S-DAYOLO achieves a 12.8% performance gain in terms of AP when only using the auxiliary domain, and a 25.0% performance gain when only using DAYOLO. These improvements demonstrate that both the auxiliary domain and DAYOLO components can effectively bridge the domain gap. By further combining these two components, the S-DAYOLO model achieves an AP of 49.3% for car detection. Similar results are obtained in the $Cityscapes \rightarrow BDD\ daytime$ scenario as well. Specifically, combining both the auxiliary domain and DAYOLO components, our method improves the Source-Only model by 5.2% (See Table V). Overall, the illustrated results present that our method can resolve the domain shift problem introduced by the different cameras used for data collection. The second and third rows in Fig. 4 show the qualitative object detection results in cross-camera adaptation.

c) Daytime-to-Nighttime Adaptation: Object detection for autonomous driving has to suffer the test of all-around-the-clock illuminations [42], [43]. The transition between daytime and nighttime poses a challenge to object detection. In this subsection, the BDD100K dataset is utilized to evaluate the adaptability of S-DAYOLO from daytime to nighttime.

The quantitative results under the daytime-to-nighttime scenario are presented in Table VI. S-DAYOLO improves the Source-Only model by 7.5% in mAP. Compared with the model that only uses DAYOLO, the model only trained by the auxiliary domain improves the performance, and its mAP is increased by 4.9%. Since the Oracle model is obtained through supervised training on the target domain, it is considered to be the upper-bound of model performance on the target domain. By further combining DAYOLO together with the auxiliary domain, the improvement pushes the performance for detection closer to the Oracle model, and the mAP achieves 34.5%. Besides, it is observed that the improvement can be well generalized to different categories, which indicates that S-DAYOLO can effectively alleviate the effects of domain shift across different categories. In general, the results demonstrate that our proposed S-DAYOLO can learn more transferable knowledge from both domains to adapt to complex environments, which is a key issue in practical AV applications. The qualitative object detection results in this adaptation task are presented in the fourth row of Fig. 4. The illustrated results show that S-DAYOLO effectively reduces the domain gap introduced by complex lighting conditions.

d) Heterogeneous Adaptation: To demonstrate whether S-DAYOLO can alleviate the influence of domain shift introduced by heterogeneous sensors, the heterogeneous domain adaptation for pedestrian detection based on multiple sensors is considered in this subsection. We build the source domain by using the annotated RGB images in the KAIST dataset and use the proposed method to detect objects in the target domain where there is lack of annotated long-wavelength infrared (LWIR) data.

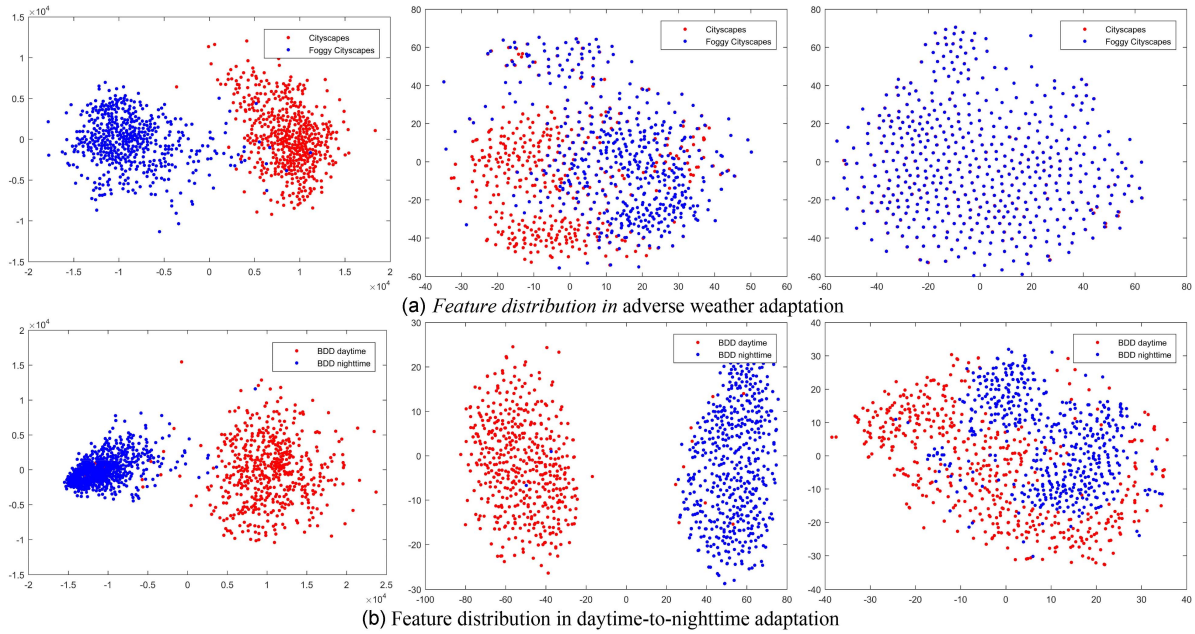
Table VII displays the adaptation results in the $KAIST\ RGB \rightarrow KAIST\ LWIR$ scenario. Compared with the Source-Only model, the proposed S-DAYOLO achieves a performance gain of 17.1% in AP for pedestrian detection, indicating that our method is able to generate more transferable features and is scalable for domain shift scenarios introduced by heterogeneous sensors. It is worth noting that all components of the proposed S-DAYOLO are designed appropriately given that the detection performance greatly drops with the removal of any of these components. The fifth row in Fig. 4 presents the qualitative object detection results.

D. Distribution Visualization

To examine whether our proposed method can bridge the gap of feature distribution extracted from the source and target samples, the t-distributed stochastic neighbor embedding (t-SNE) of feature distribution is used for visualization following the approach by van der Maaten and Hinton [44]. Fig. 5(a) and 5(b) present the t-SNE results of feature distribution in adverse weather adaptation and daytime-to-nighttime adaptation, respectively. The first column shows the feature distributions of the original data between the source and target samples. The feature distributions generated by the Source-Only model and our proposed S-DAYOLO are presented on the second and third columns, respectively. The third column results in Fig. 5 demonstrate similar feature distributions extracted by the proposed S-DAYOLO between the source and target samples, which suggests that S-DAYOLO can helpfully tackle the domain shift problem caused by artificial fog. The illustrated results in daytime-to-nighttime adaptation shows that the source-only model has a significant gap between the features generated from source and target samples, but the feature distribution interval after adaptation is small. This demonstrates that the proposed S-DAYOLO can helpfully bridge the distribution gap between source and target samples.

E. Analysis of Invariant Representations

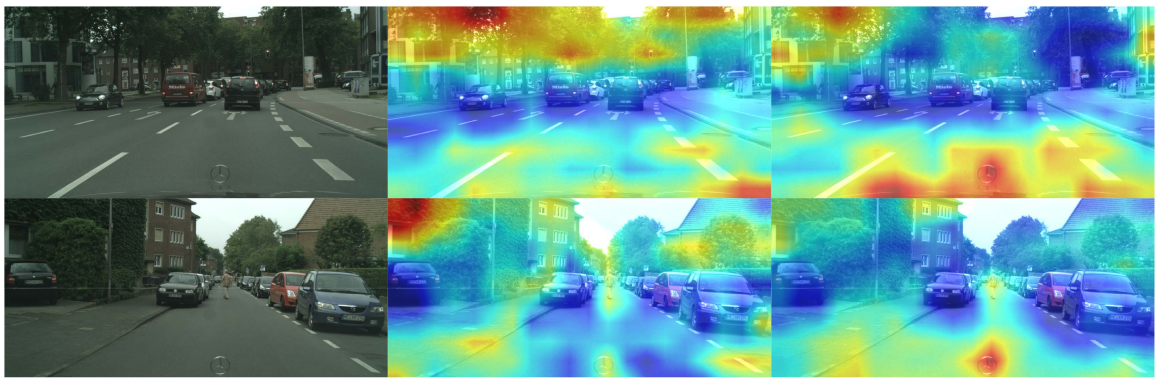
To demonstrate that the invariant representations are generated by our proposed S-DAYOLO from different domains, the feature activation maps of the SSP layer from DAYOLO are visualized in Fig. 6. It is found that the S-DAYOLO model learns



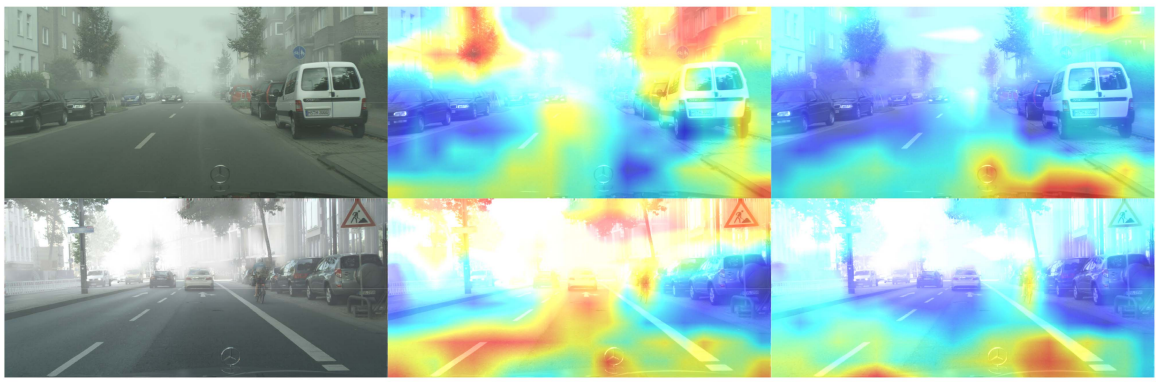
(a) *Feature distribution in* adverse weather adaptation

(b) Feature distribution in daytime-to-nighttime adaptation

Fig. 5. Visualization results of feature distribution by t-SNE. The first column is for original images. The second and third columns are for feature distribution visualization based on the Source-Only model and the S-DAYOLO, respectively. The red dots and blue dots represent the feature distributions learn from the source and target samples, respectively.



(a) Activation maps of YOLOv5s and S-DAYOLO in source domain



(b) Activation maps of YOLOv5s and S-DAYOLO model in target domain

Fig. 6. Visualization of the domain using activation map. The first column is for original images. The second and third columns are for activation maps visualization based on the Source-Only and S-DAYOLO model, respectively.

TABLE V
THE RESULTS OF MAP (%) ON THE BDD DAYTIME VALIDATION SET. THE METHODS ARE TRAINED USING SOURCE SAMPLES FROM CITYSCAPES AND TARGET SAMPLES FROM BDD DAYTIME. "ORACLE" DENOTES A YOLOV5S MODEL WITH THE SAME SETTING SUPERVISED TRAINED WITH TARGET SAMPLES

Method	Person	Rider	Car	Truck	Bus	Motor	Bike	Train	mAP
Source-Only	39.5	21.1	58.8	23.3	20.7	15.4	19.9	-1	28.4
Ours w/o DAYOLO	47.0	24.0	62.9	**29.9**	**29.2**	**17.5**	20.8	-1	33.0
Ours (DAYOLO)	42.8	25.5	62.4	28.5	23.3	15.5	**22.2**	-1	31.5
Ours (S-DAYOLO)	**48.4**	**29.1**	**64.5**	29.5	28.6	14.4	20.5	-1	**33.6**
Oracle	55.5	32.6	72.7	51.4	50.2	40.3	31.4	-1	47.7

TABLE VI
THE RESULTS OF MAP (%) ON THE BDD NIGHTTIME VALIDATION SET. THE METHODS ARE TRAINED USING SOURCE SAMPLES FROM BDD DAYTIME AND TARGET SAMPLES FROM BDD NIGHTTIME. "ORACLE" DENOTES A YOLOV5S MODEL WITH THE SAME SETTING SUPERVISED TRAINED WITH TARGET SAMPLES

Method	Person	Rider	Car	Truck	Bus	Motor	Bike	Train	mAP
Source-Only	36.9	18.8	59.4	25.9	30.1	18.2	26.4	0.0	27.0
Ours w/o DAYOLO	43.4	24.6	63.3	36.6	39.5	17.3	30.5	0.0	31.9
Ours (DAYOLO)	41.9	20.5	61.1	34.7	34.5	17.2	28.4	0.0	29.8
Ours (S-DAYOLO)	**44.8**	**25.1**	**63.9**	**39.4**	**42.6**	**27.5**	**32.5**	0.0	**34.5**
Oracle	50.7	34.6	72.4	52.6	50.3	32.2	46.1	0.0	42.3

TABLE VII
AVERAGE PRECISION (AP) RESULTS (%) OF PEDESTRIAN ON KAIST LWIR. THE METHODS ARE TRAINED USING SOURCE SAMPLES FROM KAIST RGB AND TARGET SAMPLES FROM KAIST LWIR. "ORACLE" DENOTES A YOLOV5S MODEL WITH THE SAME SETTING SUPERVISED TRAINED WITH TARGET SAMPLES

Method	Pedestrian AP
Source-Only	42.6
Ours w/o DAYOLO	47.9
Ours (DAYOLO)	54.8
Ours (S-DAYOLO)	**59.7**
Oracle	74.3

invariant representations from auxiliary and target samples, and the region of interest is activated. Fig. 6 illustrates that the YOLOv5s model trained with source samples underworks in target samples, while the S-DAYOLO model performs better in target samples. The results demonstrate that S-DAYOLO can extract effective features of interest from the target samples. In other words, the proposed S-DAYOLO trained with auxiliary samples can work well in target samples and achieve the objective of domain adaptation.

F. Comparison With the Vanilla YOLO Model

To demonstrate that our proposed method can significantly improve the detection performance of the vanilla YOLO model in cross-domain scenarios, the mAP gain results of S-DAYOLO and their comparisons with a similar work named MS-DAYOLO [27] are presented in Table VIII. Comparing with vanilla YOLOv5s, it can be found that S-DAYOLO has a significant positive mAP gain (+5.4% to +25.6%) in the examined cross-domain scenarios and achieves the best detection results in cross-camera (KITTI→Cityscapes) and heterogeneous (KAIST RGB→KAIST LWIR) adaptation scenarios. The vanilla YOLOv4 outperforms vanilla YOLOv5s in most of the cross-domain scenarios except heterogeneous adaptation. A possible reason is that the high complexity of vanilla YOLOv4, as shown in Table IX, makes it have strong generalization ability [28]. Benefiting from the strong generalization ability of vanilla YOLOv4, MS-DAYOLO achieves the highest mAP values of 42.3%, 40.2%, and 42.7% in cross-domain scenarios from clear to foggy (Cityscapes→Foggy Cityscapes), sunny to rainy (Cityscapes→BDD rainy), and cross-camera (Cityscapes→BDD daytime), respectively. However, as shown in the mAP gain results in Table VIII by comparing with the corresponding vanilla model, the mAP performance improvement of our S-DAYOLO over the vanilla YOLOv5s greatly outperforms the improvement of MS-DAYOLO over YOLOv4, which demonstrates that our proposed method can significantly improve the performance of the vanilla model. Besides, MS-DAYOLO gets a negative mAP gain (-2.3%) in daytime-to-nighttime adaptation scenarios, while the corresponding number obtained by our S-DAYOLO is +7.5%, supporting that our proposed method has stronger scalability and can adapt to more cross-domain scenarios.

To examine the complexity and efficiency of the models, the floating point operations (FLOPs), number of parameters and frame per second (FPS) of the models are shown in Table IX. Compared with other competitors like Faster RCNN [13] and Mask RCNN [45], YOLOv5s is less complex and more efficient. It is found that our proposed method only increases 2.8G FLOPs and 2.08M parameters during training when comparing with vanilla YOLOv5s. In the inference process, the domain adaptive components are removed and the vanilla YOLOv5s architecture with adaptive weights is used, which implies that our proposed method can get the same result of FPS as vanilla YOLOv5s.

G. Novelties of Our Proposed Method

Unlike the previous methods, the novelties of this work can be summarized from the following two aspects. First, a generative pipeline for building an auxiliary domain is newly proposed, which can reduce the different between source and target images at the input level. In adverse weather adaptation (Cityscapes→Foggy Cityscapes), the use of auxiliary domains achieves a 15.6% mAP gain when comparing with the Source-Only model. Second, DAYOLO with a category-consistent

TABLE VIII
THE MAP AND MAP GAIN RESULTS (%) OF DIFFERENT CROSS-DOMAIN SCENARIOS

Adaptation Scenarios		YOLOv4	MS-DAYOLO[27]	mAP Gain	YOLOv5s	Our (S-DAYOLO)	mAP Gain
Adverse weather	Cityscapes→Foggy Cityscapes	40.2	42.3	+2.1	22.5	39.0	+16.5
	Cityscapes→BDD rainy	38.6	40.2	+1.6	24.6	37.6	+13.0
Cross-camera	KITTI→Cityscapes	37.5	44.7	+7.2	23.7	49.3	+25.6
	Cityscapes→BDD daytime	40.6	42.7	+2.1	28.4	33.6	+5.4
Daytime-to-nighttime	BDD daytime→BDD nighttime	37.3	35.0	-2.3	27.0	34.5	+7.5
Heterogeneous	KAIST RGB→KAIST LWIR	33.0	42.6	+9.6	42.6	59.7	+17.1

TABLE IX
THE COMPLEXITY AND EFFICIENCY OF THE MODELS

Model	FLOPs (G)	Parameter number (M)	FPS
Faster RCNN	40.9	33.79	9.8
Mask RCNN	42.5	35.91	8.1
YOLOv4	142.4	65.72	-
YOLOv5s	16.9	7.27	113.8
Our (S-DAYOLO)	19.7	9.35	113.8

regularization module and adaptation modules for image-level and instance-level features is newly designed, which can generate domain-invariant representations to stepwise remove the gap between domains. Compared with the Source-Only model, DAYOLO gets a 25% performance gain for car detection in the cross-camera scenarios (KITTI→Cityscapes). S-DAYOLO is designed by comprehensively considering an auxiliary domain and DAYOLO, which can further learn an advanced cross-domain object detector but does not increase the computation burden in the inference stage. Compared with the Source-Only model, S-DAYOLO achieves at least +5.4% improvement of mAP in the examined domain shift scenarios.

Given these novelties and the favorable results obtained in domain shift scenarios, our proposed method is demonstrated to effectively alleviate the domain gap with a stronger adaptation capability than the state-of-the-art methods. Therefore, our developed method provides a promising solution to the cross-domain object detection problems in various weather and illumination conditions to improve the design of ADASs and AVs.

H. Limitations and Future Work

The main limitation of this study is that our proposed domain adaptation method is only examined in YOLOv5s. Although our proposed method is in essence designed for cross-domain detection tasks and it has been convinced to be effective in S-DAYOLO based on YOLOv5s, if the method can be effectively extended to SSD [15] still needs further investigation. Besides, our future work will also consider multi-source domains to improve the detection robustness in object detection tasks based on domain adaptation.

V. CONCLUSION

In this paper, a framework named S-DAYOLO is developed to address the performance degradation of object detection models in domain shift driving scenarios, which is crucial for the design of ADASs and AVs. Through the construction of auxiliary domains and the learning of domain-invariant representations, the training of cross-domain object detection is realized. Using our proposed method, a well-trained detection model can be obtained in new scenarios without using annotated labels. The detection model trained by our framework can remove the modules for domain adaptation when the model is deployed so as not to increase the time consumed in the inference phase. In the conducted experiments, our framework has been examined in various domain shift scenarios based on five public driving datasets, and our framework has shown superior object detection performance in the examined scenarios. The distribution visualization and analysis of invariant representations further explain the effectiveness of the S-DAYOLO framework in cross-domain object detection tasks.

REFERENCES

[1] Z. Chen and X. Huang, "Pedestrian detection for autonomous vehicle using multi-spectral cameras," *IEEE Trans. Intell. Veh.*, vol. 4, no. 2, pp. 211–219, Jun. 2019,.
[2] T. Gao, H. Pan, and H. Gao, "Monocular 3D object detection with sequential feature association and depth hint augmentation," *IEEE Trans. Intell. Veh.*, to be published, doi: 10.1109/TIV.2022.3143954.
[3] G. Li, Z. Ji, and X. Qu, "Stepwise domain adaptation (SDA) for object detection in autonomous vehicles using an adaptive CenterNet," *IEEE Trans. Intell. Transp. Syst.*, to be published, doi: 10.1109/TITS.2022.3164407.
[4] G. Li, Y. Yang, and X. Qu, "Deep learning approaches on pedestrian detection in hazy weather," *IEEE Trans. Ind. Electron.*, vol. 67, no. 10, pp. 8889–8899, Oct. 2020.
[5] U. Michieli, M. Biasetton, G. Agresti, and P. Zanuttigh, "Adversarial learning and self-teaching techniques for domain adaptation in semantic segmentation," *IEEE Trans. Intell. Veh.*, vol. 5, no. 3, pp. 508–518, Sep. 2020.
[6] A. Zappone, M. Di Renzo, M. Debbah, T. T. Lam, and X. Qian, "Model-aided wireless artificial intelligence: Embedding expert knowledge in deep neural networks for wireless system optimization," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 60–69, Sep. 2019.
[7] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3339–3348.
[8] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," *Adv. Neural Inf. Process. Syst.*, vol. 31, pp. 1647–1657, 2018.
[9] G. Li, Z. Ji, Y. Chang, S. Li, X. Qu, and D. Cao, "ML-ANet: A transfer learning approach using adaptation network for multi-label image classification in autonomous driving," *Chin. J. Mech. Eng.*, vol. 34, no. 1, pp. 1–11, Aug. 2021.
[10] Ultralytics, "Yolov5," 2021, Accessed: Apr. 2022. [Online]. Available: https://github.com/ultralytics/yolov5
[11] Q. Wen, Z. Luo, R. Chen, Y. Yang, and G. Li, "Deep learning approaches on defect detection in high resolution aerial images of insulators," *Sensors*, vol. 21, no. 4, pp. 1033, Feb. 2021.
[12] H. Wang, Y. Yu, Y. Cai, X. Chen, L. Chen, and Y. Li, "Soft-weighted-average ensemble vehicle detection method based on single-stage and two-stage deep learning models," *IEEE Trans. Intell. Veh.*, vol. 6, no. 1, pp. 100–109, Mar. 2021.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 1137–1149, 2017.
[14] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13 029–13 038.
[15] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[16] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 3320–3328, 2014.
[17] Y. Zhu, F. Zhuang, and D. Wang, "Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 5989–5996.
[18] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Central moment discrepancy (CMD) for domain-invariant representation learning," May 2019, *arXiv.1702.08811*. [Online]. Available: https://doi.org/10.48550/arXiv.1702.08811
[19] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.
[20] Q. Zhou, S. Wang, and Y. Xing, "Multiple adversarial networks for unsupervised domain adaptation," *Knowl.-Based Syst.*, vol. 212, Nov. 2021, Art. no. 106606. [Online]. Available: https://doi.org/10.1016/j.knosys.2020
[21] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3722–3731.
[22] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 687–696.
[23] Z. He and L. Zhang, "Multi-adversarial faster-rcnn for unrestricted object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6668–6677.
[24] G. Zhao, G. Li, R. Xu, and L. Lin, "Collaborative training between region proposal localization and classification for domain adaptive object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 86–102.
[25] P. Su *et al.*, "Adapting object detectors with conditional domain normalization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 403–419.
[26] H. Zhang, G. Luo, Y. Tian, K. Wang, H. He, and F. Y. Wang, "A virtual-real interaction approach to object instance segmentation in traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 863–875, Feb. 2021.
[27] M. Hnewa and H. Radha, "Multiscale domain adaptive YOLO for cross-domain object detection," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 3323–3327.
[28] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," Apr. 2020, *arXiv.2004.10934*. [Online]. Available: https://doi.org/10.48550/arXiv.2004.10934
[29] X. Li, Y. Wang, L. Yan, K. Wang, F. Deng, and F. Y. Wang, "ParallelEye-CS: A new dataset of synthetic images for testing the visual intelligence of intelligent vehicles," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 9619–9631, Oct. 2019.
[30] W. Zhang, K. Wang, Y. Liu, Y. Lu, and F. Y. Wang, "A parallel vision approach to scene-specific pedestrian detection," *Neurocomputing*, vol. 394, pp. 114–126, Apr. 2020.
[31] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 1, pp. 151–175, May 2010.
[32] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[33] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.
[34] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
[35] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 973–992, Sep. 2018.
[36] F. Yu *et al.*, "BDD100K: A diverse driving video database with scalable annotation tooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2636–2645.
[37] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
[38] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037–1045.
[39] G. Li, L. Yang, S. Li, X. Luo, X. Qu, and P. Green, "Human-like decision making of artificial drivers in intelligent transportation systems: An end-to-end driving behavior prediction approach," *IEEE Intell. Transp. Syst. Mag.*, to be published, doi: 10.1109/MITS.2021.3085986.
[40] G. Li, Y. Lin, and X. Qu, "An infrared and visible image fusion method based on multi-scale transformation and norm optimization," *Inf. Fusion*, vol. 71, pp. 109–129, Feb. 2021.
[41] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 49–56.
[42] M. Schutera, M. Hussein, J. Abhau, R. Mikut, and M. Reischl, "Night-to-day: Online image-to-image translation for object detection within autonomous driving by night," *IEEE Trans. Intell. Veh.*, vol. 6, no. 3, pp. 480–489, Sep. 2021.
[43] G. Li, Y. Yang, X. Qu, D. Cao, and K. Li, "A deep learning based image enhancement approach for autonomous driving at night," *Knowl.-Based Syst.*, vol. 213, Jan. 2021, Art. no. 106617.
[44] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 559–572, Nov. 2008.
[45] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

Guofa Li (Member, IEEE) received the Ph.D. degree in mechanical engineering from Tsinghua University, Beijing, China, in 2016. He is currently an Associate Research Professor with the College of Mechatronics and Control Engineering, Shenzhen University, Guangdong, China. He has authored or coauthored more than 60 papers in his research field, which include environment perception, driver behavior analysis, and human-like decision-making based on artificial intelligence technologies in autonomous vehicles and intelligent transportation systems. He was the recipient of the Young Elite Scientists Sponsorship Program in China, and best paper awards from the China Association for Science and Technology (CAST) and *Automotive Innovation*. In addition, he is an Associate Editor for IEEE SENSORS JOURNAL, and the Guest Editor of *IEEE Intelligent Transportation Systems Magazine* and *Automotive Innovation*.

Zefeng Ji received the bachelor's degree in automation in 2019 from Shenzhen University, Shenzhen, China, where he is currently working toward the master's degree in automation with the College of Mechatronics and Control Engineering. His research interests include computer vision, deep learning, and transfer learning in automotive and transportation engineering.

Xingda Qu received the Ph.D. degree in human factors and ergonomics from Virginia Tech, Blacksburg, VA, USA, in 2008. He is currently a Professor with the Institute of Human Factors and Ergonomics, Shenzhen University, Shenzhen, China. His research interests include transportation safety, occupational safety and health, and human computer interaction.

Rui Zhou received the B.Sc. degree in automobile engineering from Tongji University, Shanghai, China, in 2010, and the M.Sc. degree in automobile engineering from the Technical University of Braunschweig, Braunschweig, Germany, in 2014. He is currently the R&D Director with Waytous Inc., China. He was a Software Engineer and Test Engineer with Daimler AG., Stuttgart, Germany, and Ford-Werke GmbH., Cologne, Germany. His research interests include autonomous vehicle, test area for intelligent-connected vehicle, and functional safety.

Dongpu Cao received the Ph.D. degree from Concordia University, Montreal, QC, Canada, in 2008. He is currently a Professor with Tsinghua University, Beijing, China. He has contributed more than 200 papers and three books. His current research interests include driver cognition, automated driving, and cognitive autonomous driving. He was the recipient of the SAE Arch T. Colwell Merit Award in 2012, IEEE VTS 2020 Best Vehicular Electronics Paper Award, and six best paper awards from international conferences. Prof. Cao was the Deputy Editor-in-Chief for *IET Intelligent Transport Systems Journal*, and an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE/ASME TRANSACTIONS ON MECHATRONICS, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE/CAA JOURNAL OF AUTOMATICA SINICA, IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, and *ASME Journal of Dynamic Systems, Measurement, and Control*. Prof. Cao is an IEEE VTS Distinguished Lecturer.