Few-shot Adaptive Object Detection with Cross-Domain CutMix

Yuzuru Nakamura[1][*][†] ⓘ, Yasunori Ishii[1][*] ⓘ, Yuki Maruyama[1] ⓘ, and Takayoshi Yamashita[2] ⓘ

[1] Panasonic Holdings Corporation,
3-1-1 Yagumo-naka-machi, Moriguchi City, Osaka 570-8501, Japan
{nakamura.yuzuru,ishii.yasunori,maruyama.yuuki}@jp.panasonic.com
[2] Chubu University, 1200 Matsumoto-cho, Kasugai City, Aichi 487-8501, Japan
takayoshi@isc.chubu.ac.jp

Abstract. In object detection, data amount and cost are a trade-off, and collecting a large amount of data in a specific domain is labor-intensive. Therefore, existing large-scale datasets are used for pre-training. However, conventional transfer learning and domain adaptation cannot bridge the domain gap when the target domain differs significantly from the source domain. We propose a data synthesis method that can solve the large domain gap problem. In this method, a part of the target image is pasted onto the source image, and the position of the pasted region is aligned by utilizing the information of the object bounding box. In addition, we introduce adversarial learning to discriminate whether the original or the pasted regions. The proposed method trains on a large number of source images and a few target domain images. The proposed method achieves higher accuracy than conventional methods in a very different domain problem setting, where RGB images are the source domain, and thermal infrared images are the target domain. Similarly, the proposed method achieves higher accuracy in the cases of simulation images to real images.

Keywords: Object detection · Domain adaptation · Few-shot learning.

1 Introduction

Systems used in various environments throughout the day, such as autonomous driving and surveillance robots, are being put to practical use. These systems require high accuracy throughout the day. Infrared cameras can capture visible images even in such situations, and they can robustly detect objects. To achieve high accuracy, object detection models require a large amount of labeled training data. It is easy to use a large amount of labeled training data of RGB images [24,7,22,34,20,4]. However, most of the images are captured using an RGB camera, and a few images are captured using an infrared camera (hereinafter

---

[*] indicates equal contribution.

[†] indicates the corresponding author.

infrared images). Therefore, to achieve high accuracy, detection models need to train with a few labeled infrared images.

One of the methods for training high accuracy detection model is transfer learning using pre-trained model trained with RGB images. However, it is difficult to improve the accuracy if the domain gap between RGB and infrared images is large [30,40]. This phenomenon, called negative transfer, occurred under the large domain gap between the training images for pre-trained model and the ones for fine-tuning model. As one of the conventional methods for overcoming the domain gap, there are style transformation methods such as CycleGAN [51,33]. GAN-based style transformations can easily convert between images with similar spectra. However, the style transformation is difficult when the spectral distributions are significantly different, such as in RGB and infrared images. A method using GRL [29,2,8] is proposed as a training method for domain adaptation in the feature space. However, if the spectral distributions between the input images are significantly different, it is difficult to align the distributions of different domains because the feature distributions are extremely different. The methods for training features that interpolate between two images are as follows: Mixup [46], BC-learning [37], CutMix [45], and CutDepth [16]. These methods train features located between two images by mixing the two images or by replacing a portion of the image with the other image. These data augmentation methods synthesize features with a mixture of different domains.

We propose a few shot adaptive object detection with cross-domain CutMix. We take advantage of the fact that data augmentation can reduce the domain gap by mixing features of domains with a large domain gap. Our method enables highly accurate object detection even for a few annotated infrared images based on a pre-trained model of RGB images. We paste a part of one domain's image onto a part of another domain's image, such as CutMix, because we overcome large domain gap. Particularly, in object detection task, the size of detection targets is smaller than that of the background. Therefore, to perform domain adaptation of small detection object, we cut out the detection object and paste it onto the other domain instead of randomly cutting out the image. Even if there is a significant difference in appearance between domains, the features of the detected objects between domains are trained to be similar to each other.

Additionally, we adapt the domain using feature-based adversarial learning. In conventional methods, the discriminator of the domain identification label does not change during training. However, the domain identification label also needs to change the label according to the pasted area because our CutMix-based method changes the pasting area during training. The conventional domain identification label cannot be used by pasting an image of another domain. Therefore, the domain identification label should be the same domain label as the input image to which the image of another domain is pasted. Since the correct domain label can be assigned according to the pasting position of the object, feature-level domain alignment can be performed even when the input image is changed, as in the proposed method.

Our contributions are as follows: we propose a few shot adaptive object detection with cross-domain CutMix so that we can adapt the domain, which looks significantly different. Furthermore, we propose an input image synthesizing method based on CutMix for cross-domains and domain identification label in discriminator for that. Through experiments, we show the effectiveness of the proposed method using RGB images as a pre-trained model and data from multiple domains such as RGB and thermal infrared images.

2   Related Work

2.1   Object Detection for Each Domain

Most object detection methods have been studied for RGB images [26,18,41]. These methods can be roughly divided into two-stage and one-stage detection methods. R-CNN and its extended technologies [11,10,32] represent the two-stage methods. YOLO [31], SSD (Single Shot Multi-box Detector) [27], and their extended technologies represent the one-stage methods. Additionally, object detection techniques that use transformer have been proposed [6,3,52,44,28], and they are expected that it will be applied to various environments.

There is research on applications in the real environment such as robots [19,36], drones [13,42], object detection for in-vehicle cameras [23,49], and license plate detection [43]. Many datasets [24,7,22,34,20,4] that can be used to train object detection are available to the public. The night scenes on these datasets are fewer than daytime scenes. Furthermore, visibility is poor because the pixel value of the subject in the RGB image captured at night is small. Thus, it is difficult to achieve high detection performance using RGB images both day and night.

Highly accurate object detection with in-vehicle cameras and outdoor drones is required for both day and night. Some methods for detecting objects using spectra information other than RGB images were proposed to detect objects with high accuracy on both day and night. Lu et al. [47] proposed object detection using RGB and infrared images in the framework of weakly supervised learning. This method focuses on the use of multispectral information, and it detects objects using a roughly aligned RGB image and an infrared image. Liu et al. [25] and Konig et al. [21] proposed methods for inputting RGB and thermal infrared images into a deep learning model and fusing their features. Highly accurate object detection is possible under various lighting conditions using RGB and infrared images simultaneously. These methods are algorithms that assume that there are numerous RGB and infrared images. Thus, they can be used if the RGB and the infrared images can be photographed in large quantities and annotated in the same environment as the inference scene. However, the cost of collecting data and annotating in each application environment is high in reality.

2.2   Knowledge Transfer to Different Domain

To reduce the cost of preparing infrared images and training high accurate object detection models, domain adaptation and transfer learning use a pre-trained model with a small number of labeled infrared and RGB images.

Akkaya et al. [1] proposed unsupervised domain adaptation between a model taken from numerous RGB and thermal infrared images for image classification. Vibashan et al. [38] used paired RGB and thermal infrared images to perform domain adaptation for object detection. When only the recognized object is displayed as in the image classification, the domains of both the difference in the sensor and shooting scene can be applied. However, when the background area without objects occupies a large area in object detection, it is difficult to adapt the domain of both the sensor difference and the shooting scene difference without using a pair of datasets. Thus, it is still a problem to adapt between different sensors and scenes for object detection.

There is a knowledge transfer method, which is by fine-tuning infrared images, using a model trained with RGB images as pre-trained model. For fine-tuning to be effective, the feature of training data between source domain and target domain must be similar. RGB and infrared images have extremely different spectrum to be imaged, and they look very different even if they have the same object and color. Negative transfer [30,40] occurs because of the difference in the distribution of these data. Therefore, knowledge transfer using a small number of labeled infrared images is difficult. In both domain adaptation and fine-tuning, the key to improving performance is transferring knowledge while making the differences between domains closer.

3   Proposed Method

3.1   Overview

In this paper, we propose high-accuracy object detection on infrared images using a large number of labeled RGB images and a small number of labeled infrared images. RGB and infrared images receive different spectra; thus, there is a significant difference in appearance, which is a large gap between domains. Therefore, we not only align the gaps between domains at the feature level using methods such as adversarial learning but also explicitly reduce the gaps between domains at the image level. This improves the accuracy of domain adaptation by converting the input image to conditions that make it adapt the domain easier.

We propose Object aware Cross-Domain CutMix (OCDC) and OCDC-based Discriminator Label (OCDCDL) based on the domain for each location. Figure 1 shows our proposed framework. We explain the outline of the proposed method using the model of the domain adaptation method based on adversarial learning proposed by Han-Kai et al. [14] as an example. This method trais using the loss of both object detection and adversarial learning to reduce the difference between domains. This proposed method is simple and easy to incorporate into the type of domain adaptation that uses adversarial learning in object detection problems, which uses few labeled images.

OCDC (Fig. 1 (a)) is a method for cutting out an object area and pasting a part of the image between domains to reduce the gap between the source and the target domains. Zhou et al. showed that there is a domain generalization
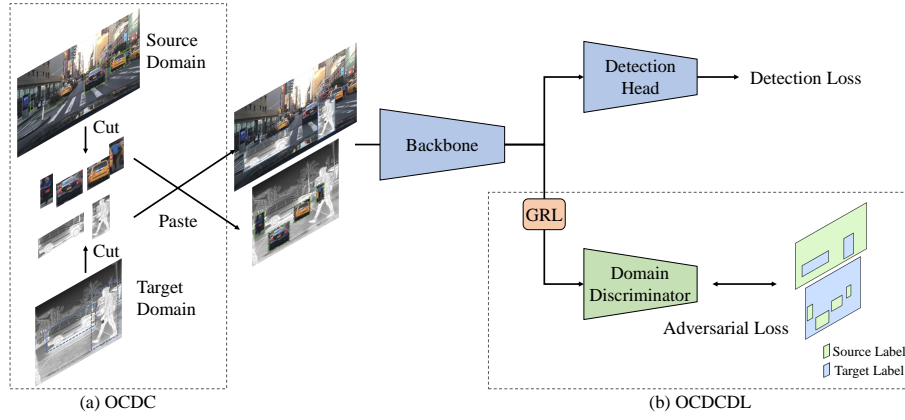
Fig. 1: The framework of our proposed method; (a) OCDC mitigates the image-level domain gap by cutting out the object area and pasting them in each other domain. (b) We adaptively determine the domain identification label based on the area pasted by the OCDC. The pasted object, which serve areas as new ground truth, are input into the detection network, and the detection and the adversarial losses are calculated.

effect by mixing images with different domains in a batch [50]. Inspired by that study, we focused on mixing object units, which is important for object detection. When the entire image is mixed, it is trained to reduce the distance between the background domains that occupy most of the image. In the proposed method, the distance between domains is emphasized rather than the background features, so the object detection performance is expected to improve.

OCDCDL (Fig. 1 (b)) is a method for adaptively changing the domain label of the discriminator based on the pasting position of other domain images. By converting the input image using OCDC, the domain identification label is no longer one value. The output feature using the input image includes information from multiple domains by pasting an image of another domain on the input image using OCDC. Thus, the conventional single identification label cannot correctly discriminate the domain. By adaptively changing the label based on the OCDC, the discriminator makes it possible to discriminate the domain even if there is information on different domains in the image.

3.2 Object aware Cross-Domain CutMix (OCDC)

For domain adaptation to data with a large gap between domains such as RGB and infrared images, we propose a method of aligning domains at the image level in addition to the conventional method of aligning domains using features. A method based on data augmentation called DomainMix augmentation [39] was proposed: as a method for reducing the domain gap at the image level. This

method reduces the difference in appearance between images by simply connecting different domains without using a deep generative model. Mixup [46], BC-learning [37], CutMix [45], and CutDepth [16] are data augmentation methods to be mixed at the image level. These studies mention that mixing images can generate features that interpolate two images. By mixing them at the image level, the features between the domains in the input image can be brought closer to each other.

However, the background area without objects occupies most of the image, and the objects that should align the domains are only part of the image. Thus, in Mixup or CutMix, which uses the entire image or mixes random regions, respectively, the gap in the background domain is small, but the gap in the domain of the object to be detected is not always small. Therefore, we propose an OCDC to cut out the object area existing in one domain and paste it onto the other domain. The domain gap is reduced at the image level by cutting out the object area from the source and target and pasting them together.

The pasting process is performed for each training iteration. The images of the source domain and that of the target domain included in the batch are used. The object to be detected is cut out from those images based on the ground truth coordinates. The object image cut out from the image of the source domain is pasted onto the image of the target domain selected at random. On the other hand, the image cut out from the object image of the target domain is pasted onto the image of the source domain selected at random. The ground truth labels for detection loss are updated by adding the ground truth label of the pasted image of another domain to that of the image where the image of another domain is pasted.

If the object images are pasted while the objects already exist, the originally existing objects will be hidden. The loss of object detection becomes large because it is difficult to detect such hidden objects. Therefore, we decide the pasting position of the image based on the overlap between the image to be pasted and the object of the image to which the object image is pasted. If the area where the object is hidden increases more than a certain percentage after pasting the image, the pasting position is reselected when deciding the pasting position. This prevents the original object from being hidden after pasting.

Additionally, to train the position and label of an object, the object detection model pastes an object image to a real-world location. For example, if the domain is adapted between the images of the in-vehicle camera, which is installed at almost the same position, the coordinates before and after pasting do not change significantly. Alternatively, the object image is pasted at the same position as the position before pasting or at a slightly shifted position.

The domain adaptation for object detection that has been used so far has insufficient consideration for object detection. On the other hand, our method using cutting out the position of objects is a new point of view that the domain gap of the object to be recognized can be reduced. In addition, we argue that it causes problems in adversarial learning and propose a solution. In the following subsections, each proposed method is explained concretely.

3.3   OCDC-based Discriminator Label (OCDCDL)

In domain adaptation using adversarial learning such as a method [14], features calculated from the source or target domains are input into the discriminator. In our proposed OCDC, the information of another domain is included in a part of the feature because the image of another domain is pasted on the part of the input image. Therefore, when one value is used for the domain identification label of the discriminator as in the conventional method, the loss of the area where the image of another domain is pasted cannot be calculated correctly.

The discriminator is trained so that neither of the two domains can be discriminated. We append the discriminator D after the backbone F. The input image of the source domain and the target domain are I_S and I_T, respectively. D outptus a domain prediction map of each pixel $D(\cdot)_{h,w}$. The source domain identification label and target domain identification label are $d = 0$ and $d = 1$, respectively. Eq. 1 is a adversarial loss $\mathcal{L}_{adv}$ and Eq. 2 is an overall loss $\mathcal{L}$.

$$\mathcal{L}_{adv}(F(I)) = -\Sigma_{h,w}[d \log D(F(I))_{h,w} + (1-d) \log(1 - D(F(I))_{h,w})], \quad (1)$$

$$\min_F \max_D \mathcal{L}(I_S, I_T) = \mathcal{L}_{det}(I_S) + \mathcal{L}_{det}(I_T) + \lambda_{adv}[\mathcal{L}_{adv}(F(I_S)) + \mathcal{L}_{adv}(F(I_T))], \quad (2)$$

where $\mathcal{L}_{det}(\cdot)$ is the detection loss, and λ_{adv} is a weight that determines the loss balance. We set λ_{disc} is 0.1 in our experiments.

However, because the features near the boundary of the area where the object of another domain is pasted using OCDC are a mixture of the features of the two domains, it is difficult to distinguish which domain is near this boundary. Thus, the loss near the boundary of the trained discriminator is smaller than that in other regions. However, information from different domains is rarely mixed at a position far from the boundary of the pasted object. Therefore, the loss of discriminator in the feature is large in that area.

Furthermore, we adaptively determine the domain identification label based on the position and domain of the image pasted by OCDC. After the pasting process, the RGB domain label is replaced with the domain identification label corresponding to the infrared image region, and the infrared domain label is replaced with the domain identification label corresponding to the RGB image region.

4   Experiments

We evaluate the effectiveness of the proposed methods using RGB images [20,5] and thermal infrared images [12,15] with a large domain gap. In these experiments, there are differences in both the spectrum and the captured scene. We compare the performance with a large amount of RGB images and a small amount of thermal infrared images. All datasets are labeled. Additionally, to verify the generalization performance of the proposed method, we evaluate the performance using real images [4] and simulation images [17].

4.1   Comparison Methods

We explain the comparison method used for each experimental setting. In our experiment, images of the target domain are used for evaluation. Source-only and target-only labels in the tables of experimental results show the evaluation results when the model is trained using only the image of the source or target domains, respectively. The fine-tuning label shows the results using a pre-trained model and fine-tuning using target domain data. The target samples label shows the number of target domain data using fine-tuning. The Domain-Adversarial Training of Neural Networks (DANN) [9], one of the adversarial learning methods, is used as the baseline of the adversarial learning method. DANN label shows the detection results using domain adaptation with DANN. We use Faster R-CNN [32] as the detection network and VGG16 [35] for the backbone. In domain adaptation, the model parameters are pre-trained in the source domain. The height size is 600 of the image resolution, but if the maximum width size is more than 1,000, we set it to 1,000 while maintaining the aspect ratio. Ours label shows the result using the proposed method, which uses the OCDC and OCDCDL. The optimizer is SGD; the learning rate, the weight decay, and the momentum are 0.001, 0.0005, and 0.9, respectively. The batch size is one. The evaluation metrics are the average precision at an intersection over union (IoU), which threshold is 0.5. The front of the arrow indicates the source domain, and the tip of the arrow indicates the target domain. The results of each experiment are shown in Table 1.

BDD100k → FLIR: The BDD100k dataset [20] is collected based on six types of weather conditions, six different scenes, and three categories of time of data; the number of images is 100,000. This dataset is annotated in ten categories. FLIR ADAS dataset [12] is an image captured by a FLIR Tau2 camera, and the number of images is 10,228. Only thermal infrared images from this dataset are used. In our experiment, the training data includes 36,728 images labeled as daytime from the BDD100k dataset as the source domain data and 8,862 thermal infrared images used as training splits from the FLIR ADAS dataset. The categories person, bicycle, and car, which are common categories.

The detection accuracy of source-only is the lowest because this does not use knowledge of the target domain. In target-only, mAP is 72.1 % using all target data named Full. However, performances considerably deteriorates when the amount of data decreased. In fine-tuning, performances are higher than the performances of target-only because fine-tuning models had a knowledge that performance improves somewhat even in different domains. The performances of DANN are higher than fine-tuning because of the effects of domain adaptation. There is no significant performance degradation due to negative transfer, but the effect of domain adaptation can be confirmed. The performances of the proposed method tend to improve overall, but in particular, the performance of the person label outperformed DANN.

When the target samples are Full, the number of each object is large, so even if the object areas are small, there are enough numbers to improve the performance by domain adaptation. However, in the case of Full, there is only

Table 1: Results on BDD100k → FLIR, Caltech → KAIST, and SIM10K → Cityscapes

| Method | Target Samples | FLIR | | | | KAIST | Cityscapes |
		Person	Bicycle	Car	mAP	Person	Car
Source-only	—	39.9	24.9	68.2	44.4	2.8	43.1
Target-only	Full	74.2	57.9	84.1	72.1	67.0	60.3
	1/2	71.5	56.0	82.3	69.9	67.6	58.1
	1/4	66.7	48.7	78.5	64.6	63.0	54.3
	1/8	61.4	41.4	75.4	59.4	58.7	52.0
	1/16	57.0	42.5	71.8	57.1	57.1	48.3
	1/32	51.3	34.9	67.3	51.2	51.3	45.2
	1/64	44.4	32.0	63.6	46.7	47.4	41.1
Fine-tuning	Full	75.0	60.5	86.3	73.9	63.4	58.1
	1/2	74.8	58.3	86.1	73.1	63.4	58.1
	1/4	72.4	53.1	85.4	70.1	63.5	57.8
	1/8	69.1	47.8	82.9	66.6	63.2	56.8
	1/16	64.6	45.4	79.5	63.2	61.7	55.7
	1/32	65.9	46.7	**83.3**	65.3	**59.5**	52.5
	1/64	64.5	41.2	82.0	62.6	**57.1**	49.8
DANN	Full	**78.1**	**63.8**	**87.0**	**76.3**	**69.4**	62.0
	1/2	77.6	**63.1**	**87.2**	76.0	71.7	61.0
	1/4	75.2	56.5	86.2	72.6	70.4	58.6
	1/8	72.4	58.8	84.5	71.9	69.1	**59.6**
	1/16	70.6	55.8	83.8	70.1	66.9	55.2
	1/32	69.4	**53.8**	82.3	68.5	57.0	54.1
	1/64	67.7	**51.8**	81.9	67.1	54.5	52.3
Ours	Full	77.8	63.5	86.9	76.1	68.4	**63.6**
	1/2	**78.3**	62.6	**87.2**	**76.1**	**73.3**	**61.8**
	1/4	**76.9**	**59.9**	**86.9**	**74.5**	**72.4**	**60.0**
	1/8	**75.4**	**60.9**	**85.7**	**74.0**	**69.7**	**59.6**
	1/16	**72.2**	**57.9**	**84.5**	**71.5**	**67.5**	**57.4**
	1/32	**71.1**	**53.8**	82.0	**69.3**	59.1	**56.5**
	1/64	**68.5**	51.6	**82.3**	**67.5**	**57.1**	**54.2**

a difference of 0.3 points, so the conventional domain adaptation does not have a difference drastically. In particular, in this experimental result, it should be noted that the target samples, which are our targets, are smaller than 1/2, rather than the detection result of Full. Under these conditions, the our method has higher performance than the conventional method in almost all cases. We confirmed about 4 point improvements over DANN even if there is a few data. For example, if we consider person label, this is because the percentage of people in the dataset was high, and the overall percentage of pasting person's images onto another domain based on the input image was high.

Caltech → KAIST: The Caltech Pedestrian dataset [5] is a dataset that contains labeled images of pedestrians captured using an in-vehicle camera. 42,782 images from this dataset are used for the images. The KAIST Multi-

spectral Pedestrian dataset [15] is a dataset captured using the FLIR A35 microbolometer LWIR camera and contains 95,000 images labeled for pedestrians. The thermal infrared image of this dataset is used. In KAIST, 7,688 thermal infrared images are used for training, and 2,252 thermal infrared images are used for testing. This is based on the procedure in the paper [48]. In this experiment, only person is used as the category.

Source-only detection accuracy is extremely low because the appearance of RGB and thermal infrared images differs significantly. Target-only and fine-tuning detection accuracy deteriorated as the target sample decreased, as demonstrated in the FLIR experiment. Particularly, fine-tuning detection accuracy was lower than target-only detection accuracy by pre-training in the source domain. However, the performance degradation is suppressed when the target samples decrease. DANN and the proposed method perform better than target-only when there are many target samples. Performance is higher than fine-tuning when the number of samples decreased. The proposed method showed more than two points higher performance than DANN when target samples are 1/32 and 1/64. This experiment shows that the proposed method is more effective in single-class object detection than in multi-class object detection. In single-class object detection, when an image of another domain is pasted on an image of another domain, the image is rarely pasted to the detected object. Therefore, few occlusion problems due to pasting occurred in the BDD100k $\rightarrow$ FLIR experiment. Since the proposed method has a remarkable effect when the number of targets is small, it is expected to be effective in applications that are often used, such as pedestrian detection.

SIM10K $\rightarrow$ Cityscapes: The SIM10K dataset [17] is a composite of 10,000 images generated by the Grand Theft Auto (GTA) game engine and is annotated with cars and other similar images. The Cityscapes dataset [4] consists of real images captured in multiple urban areas and segmentation labels. We used the circumscribed rectangle of the object segmentation label as the bounding box for evaluation in the car categories. Furthermore, we used 10,000 composite images from SIM10K as a training set. Note that 2,975 images, which are training splits from Cityscapes, are used as training data, and 500 images, which are validation splits, are used as evaluation data. The evaluation is performed using the common category of car.

Similar to the previous results, the source-only detection accuracy is the lowest. If the target domain detection accuracy has a large amount of data, the target-only detection performance is high, but when the amount of data is small, the performance is significantly reduced. Fine-tuning detection accuracy reduces performance degradation when the amount of data is small. However, when the amount of data is small, the performance of DANN improves, so the effect of domain adaptation can be confirmed. When the target samples were 1/8, the proposed method and DANN had the same accuracy. Cityscapes are in-vehicle camera images, and the size of the car in the image is much larger than that of a person. Therefore, even if the number of data is reduced a little, the vehicle area can be used for domain adaptation in the same area as the background, so the

accuracy is not so decrease. In this experiment, we confirmed that the proposed method is effective not only for domain adaptation between RGB and infrared images but also for conventional problem settings. This experiment showed that the proposed method is a general-purpose technique that can be used in various source and target data domains.

4.2 Ablation Study

This section shows the comparison results under different experimental conditions. In either case, a comparison is made under the evaluation conditions of BDD → FLIR.

Contribution of Components: We evaluate the effect of OCDC and OCD-CDL. Table 2 shows the results. For Person, the method using both OCDC and OCDCDL had high performance. On the other hand, Bicycle and Car have high performance even with OCDC alone. Our experiment do not consider labels near the boundaries of objects. Thus, even if a person with a small area makes a mistake in the discriminator label near the object boundary, the effect on detection accuracy is small. However, Bicycle and Car have a large area. Therefore, the performance decrease if the discriminator label near the object boundary is mistaken. However, in comparison with mAP, our proposed methods have the highest accuracy and effectiveness.

Region Selection Strategies: We compare the accuracy of whether the pasting position and scale are the same before and after pasting in OCDC. Table 3 shows the experimental results. A fixed label indicates position or scale are the same before and after pasting, and a random label indicates that the position and scale are set randomly. The detection accuracy in many cases is higher if the same position is maintained before and after pasting. For example, in the case of an in-vehicle camera, objects are concentrated on the lower side of the image. The object detection model trains a set of the position and the class. To train the relationship between the position and the class, which is unlikely to occur, does not have a positive effect on the inference result. In our experiment, by making the pasting position and size the same, the detection model is able to train the positions and scales that are likely to occur during inference.

4.3 Qualitative Results

Figure 2 is object detection results of FLIR, KAIST, and Cityscapes after domain adaptation performed in subsection 4.1. At the top of each dataset is the result when target samples is 1/16, and at the bottom of each dataset is the result when it is 1/64. The comparison methods are (a) fine-tuning, (b) DANN, and (c) Ours, respectively, and (d) Ground Truth. The car detection result is shown in magenta, and the person detection result is shown in cyan.

In the FLIR results, there is no difference in the car detection results, but there is a difference in the person detection results. In fine-tuning and DANN, even people with similar reflection intensities in thermal infrared images are not detected. On the other hand, the proposed method detects objects with similar

Table 2: Results on contribution of components

Target Samples	OCDC	OCDCDL	Person	Bicycle	Car	mAP
Full			**78.1**	**63.8**	**87.0**	**76.3**
	✓		77.8	63.2	86.9	76.0
	✓	✓	77.8	63.5	86.9	76.1
1/2			77.6	63.1	87.2	76.0
	✓		78.2	**63.2**	**87.4**	**76.3**
	✓	✓	**78.3**	62.6	87.2	76.1
1/4			75.2	56.5	86.2	72.6
	✓		76.7	**61.9**	86.8	**75.1**
	✓	✓	**76.9**	59.9	**86.9**	74.5
1/8			72.4	58.8	84.5	71.9
	✓		74.4	58.5	85.5	72.8
	✓	✓	**75.4**	**60.9**	**85.7**	**74.0**
1/16			70.6	55.8	83.8	70.1
	✓		72.1	54.9	**84.9**	70.6
	✓	✓	**72.2**	**57.9**	84.5	**71.5**
1/32			69.4	53.8	82.3	68.5
	✓		71.0	**54.0**	**82.6**	69.2
	✓	✓	**71.1**	53.8	82.0	**69.3**
1/64			67.7	51.8	81.9	67.1
	✓		68.1	**52.5**	81.8	**67.5**
	✓	✓	**68.5**	51.6	**82.3**	**67.5**

reflection intensities, even if they are people far away. The proposed method adapted the domain to information about the reflection intensity of persons, which is a small area in the image. On the other hand, the conventional methods did not fully adapt the domain, so some objects could not be detected.

In the KAIST results, the conventional methods did not detect some small persons. In domain adaptation, it is difficult to adapt information on small objects because even if the information in a small object is ignored, the loss of the detection model decreases. However, the proposed method makes it easier to detect even small objects by explicitly giving information from other domains to the input image.

In the Cityscapes results, in both 1/16 and 1/64, the farthest car on the left side was not detected by conventional methods, but the proposed method could detect the car. This is because it is difficult to adapt the domain of a small object, which is the same reason as in the case of KAIST. In this experiment, we clarified the importance of explicitly giving information of other domains to the input image in domain adaptation of small objects as in the proposed method.

Table 3: Results on region selection strategies

Target Samples	Position	Scaling	Person	Bicycle	Car	mAP
Full	Fixed	Fixed	77.8	**63.5**	86.9	**76.1**
	Fixed	Random	**77.9**	62.8	**87.0**	75.9
	Random	Fixed	76.0	62.5	86.5	75.0
	Random	Random	76.4	62.2	86.6	75.0
1/2	Fixed	Fixed	**78.3**	62.6	87.2	76.1
	Fixed	Random	**78.3**	**63.4**	**87.4**	**76.4**
	Random	Fixed	77.2	62.5	87.0	75.5
	Random	Random	77.0	61.2	87.0	75.1
1/4	Fixed	Fixed	76.9	59.9	**86.9**	74.5
	Fixed	Random	**77.3**	**61.4**	**86.9**	**75.2**
	Random	Fixed	75.9	59.8	86.3	74.0
	Random	Random	76.1	59.4	86.3	73.9
1/8	Fixed	Fixed	**75.4**	**60.9**	**85.7**	**74.0**
	Fixed	Random	74.6	58.1	85.5	72.7
	Random	Fixed	73.4	58.6	85.2	72.4
	Random	Random	73.8	58.0	85.1	72.3
1/16	Fixed	Fixed	**72.2**	**57.9**	**84.5**	**71.5**
	Fixed	Random	72.1	56.4	84.4	71.0
	Random	Fixed	71.3	53.9	84.1	69.8
	Random	Random	70.7	56.3	84.0	70.4
1/32	Fixed	Fixed	**71.1**	**53.8**	82.0	**69.3**
	Fixed	Random	70.2	52.0	**82.9**	68.4
	Random	Fixed	70.5	53.4	82.3	68.7
	Random	Random	69.2	51.9	82.4	67.8
1/64	Fixed	Fixed	**68.5**	**51.6**	**82.3**	**67.5**
	Fixed	Random	68.4	51.1	81.9	67.1
	Random	Fixed	67.7	46.9	81.8	65.5
	Random	Random	66.8	51.1	82.1	66.6

5 Conclusion

We proposed few-shot supervised domain adaptation for object detection in cases with large domain gaps, such as RGB and thermal infrared images. Although the number of infrared images is smaller than that of RGB images, the performance is improved compared with the conventional domain identification via OCDC method for reducing the gap between domains for the input image we proposed and the corresponding change in the domain identification label of the discriminator (OCDCDL). Furthermore, it was confirmed that the proposed method is effective by a comparative experiments in which the number of data was changed, and the versatility of the method was shown by a comparative experiments using various data.

BDD100k → FLIR
Target Samples: 1/16

Target Samples: 1/64

Caltech → KAIST
Target Samples: 1/16

Target Samples: 1/64

SIM10K → Cityscapes
Target Samples: 1/16

Target Samples: 1/64

(a) Fine-tuning   (b) DANN   (c) Ours   (d) Ground Truth
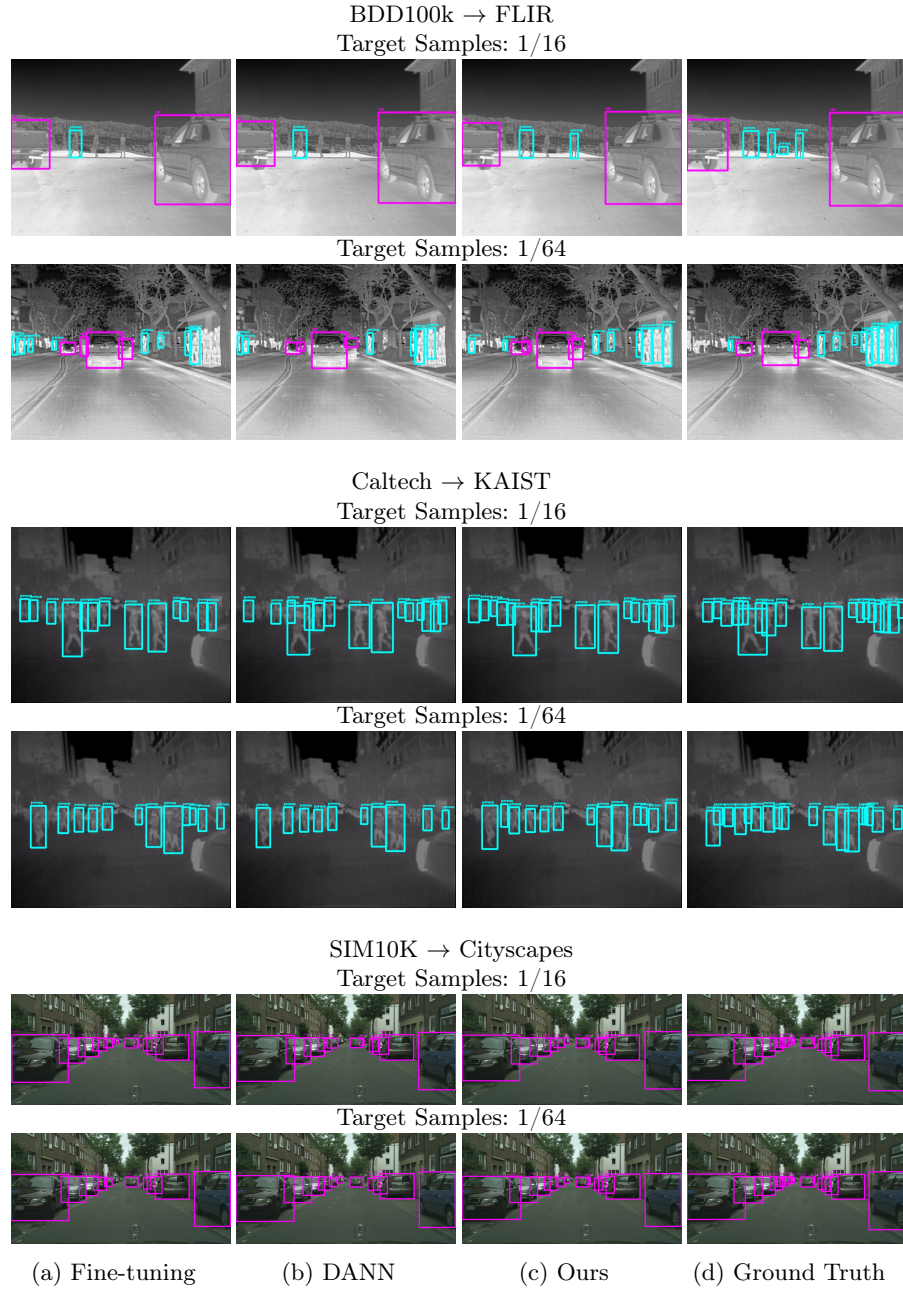
Fig. 2: There are the detection results using (a) Fine-tuning, (b) DANN, (c) our method, and (d) ground truth, respectively. The bounding box colored cyan indicates person, and the bounding box colored magenta indicates car, respectively.

References

1. Akkaya, I.B., Altinel, F., Halici, U.: Self-training guided adversarial domain adaptation for thermal imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4317–4326 (2021). https://doi.org/10.1109/CVPRW53098.2021.00488

2. Bolte, J.A., Kamp, M., Breuer, A., Homoceanu, S., Schlicht, P., Hüger, F., Lipinski, D., Fingscheidt, T.: Unsupervised domain adaptation to improve image segmentation quality both in the source and target domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1404–1413 (2019). https://doi.org/10.1109/CVPRW.2019.00181

3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 213–229 (2020). https://doi.org/10.1007/978-3-030-58452-8_13

4. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3213–3223 (2016). https://doi.org/10.1109/CVPR.2016.350

5. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 304–311 (2009). https://doi.org/10.1109/CVPR.2009.5206631

6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020). https://doi.org/10.48550/arXiv.2010.11929

7. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision (IJCV) **111**(1), 98–136 (2015). https://doi.org/10.1007/s11263-014-0733-5

8. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: Proceedings of the International Conference on Machine Learning (ICML). vol. 37, pp. 1180–1189 (2015). https://doi.org/10.48550/arXiv.1409.7495

9. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59), 1–35 (2016). https://doi.org/10.48550/arXiv.1505.07818

10. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1440–1448 (2015). https://doi.org/10.1109/ICCV.2015.169

11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 580–587 (2014). https://doi.org/10.1109/CVPR.2014.81

12. Group, F., et al.: Flir thermal dataset for algorithm training. URL: https://www. flir. co. uk/oem/adas/adas-dataset-form/(May 2019) (2018)

13. Hossain, S., Lee, D.j.: Deep learning-based real-time multiple-object detection and tracking from aerial imagery via a flying robot with gpu-based embedded devices. Sensors **19**(15), 3371 (2019). https://doi.org/10.3390/s19153371

14. Hsu, H.K., Yao, C.H., Tsai, Y.H., Hung, W.C., Tseng, H.Y., Singh, M., Yang, M.H.: Progressive domain adaptation for object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 738–746 (2020). https://doi.org/10.1109/WACV45572.2020.9093358

15. Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1037–1045 (2015). https://doi.org/10.1109/CVPR.2015.7298706

16. Ishii, Y., Yamashita, T.: Cutdepth: Edge-aware data augmentation in depth estimation. arXiv preprint arXiv:2107.07684 (2021). https://doi.org/10.48550/arXiv.2107.07684

17. Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 746–753 (2017). https://doi.org/10.1109/ICRA.2017.7989092

18. Kang, J., Tariq, S., Oh, H., Woo, S.S.: A survey of deep learning-based object detection methods and datasets for overhead imagery. IEEE Access **10**, 20118–20134 (2022). https://doi.org/10.1109/ACCESS.2022.3149052

19. Khokhlov, I., Davydenko, E., Osokin, I., Ryakin, I., Babaev, A., Litvinenko, V., Gorbachev, R.: Tiny-yolo object detection supplemented with geometrical data. In: Proceedings of the IEEE Vehicular Technology Conference (VTC). pp. 1–5 (2020). https://doi.org/10.1109/VTC2020-Spring48590.2020.9128749

20. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 577–593 (2018). https://doi.org/10.1007/978-3-030-01216-8_35

21. Konig, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., Teutsch, M.: Fully convolutional region proposal networks for multispectral person detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 243–250 (2017). https://doi.org/10.1109/CVPRW.2017.36

22. Krasin, I., Duerig, T., Alldrin, N., Veit, A., Abu-El-Haija, S., Belongie, S., Cai, D., Feng, Z., Ferrari, V., Gomes, V., Gupta, A., Narayanan, D., Sun, C., Chechik, G., Murphy, K.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github.com/openimages (2016)

23. Lin, C., Lu, J., Wang, G., Zhou, J.: Graininess-aware deep feature learning for pedestrian detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 745–761 (2018). https://doi.org/10.1007/978-3-030-01240-3_45

24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48

25. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. arXiv preprint arXiv:1611.02644 (2016). https://doi.org/10.48550/arXiv.1611.02644

26. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: A survey. International Journal of Computer Vision (IJCV) **128**(2), 261–318 (2020). https://doi.org/10.1007/s11263-019-01247-4

27. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 21–37 (2016). https://doi.org/10.1007/978-3-319-46448-0$_2$

28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9992–10002 (2021). https://doi.org/10.1109/ICCV48922.2021.00986

29. Osumi, K., Yamashita, T., Fujiyoshi, H.: Domain adaptation using a gradient reversal layer with instance weighting. In: Proceedings of the International Conference on Machine Vision Applications (MVA). pp. 1–5 (2019). https://doi.org/10.23919/MVA.2019.8757975

30. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering (TKDE) **22**(10), 1345–1359 (2009). https://doi.org/10.1109/TKDE.2009.191

31. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91

32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). vol. 28 (2015). https://doi.org/10.48550/arXiv.1506.01497

33. Shang, Q., Hu, L., Li, Q., Long, W., Jiang, L.: A survey of research on image style transfer based on deep learning. In: Proceedings of the International Conference on Artificial Intelligence and Advanced Manufacture (AIAM). pp. 386–391 (2021). https://doi.org/10.1109/AIAM54119.2021.00084

34. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8429–8438 (2019). https://doi.org/10.1109/ICCV.2019.00852

35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015). https://doi.org/10.48550/arXiv.1409.1556

36. Szemenyei, M., Estivill-Castro, V.: Fully neural object detection solutions for robot soccer. Neural Computing and Applications pp. 1–14 (2021). https://doi.org/10.1007/s00521-021-05972-1

37. Tokozume, Y., Ushiku, Y., Harada, T.: Between-class learning for image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5486–5494 (2018). https://doi.org/10.1109/CVPR.2018.00575

38. Vs, V., Poster, D., You, S., Hu, S., Patel, V.M.: Meta-uda: Unsupervised domain adaptive thermal object detection using meta-learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3697–3706 (2022). https://doi.org/10.1109/WACV51458.2022.00375

39. Wang, W., Liao, S., Zhao, F., Kang, C., Shao, L.: Domainmix: Learning generalizable person re-identification without human annotations. In: Proceedings of the British Machine Vision Conference (BMVC) (2021). https://doi.org/10.48550/arXiv.2011.11953
40. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big Data **3**(9) (2016). https://doi.org/10.1186/s40537-016-0043-6
41. Wu, X., Sahoo, D., Hoi, S.C.: Recent advances in deep learning for object detection. Neurocomputing **396**, 39–64 (2020). https://doi.org/10.1016/j.neucom.2020.01.085
42. Wu, Z., Suresh, K., Narayanan, P., Xu, H., Kwon, H., Wang, Z.: Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1201–1210 (2019). https://doi.org/10.1109/ICCV.2019.00129
43. Xu, Z., Yang, W., Meng, A., Lu, N., Huang, H., Ying, C., Huang, L.: Towards end-to-end license plate detection and recognition: A large dataset and baseline. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 261–277 (2018). https://doi.org/10.1007/978-3-030-01261-8_16
44. Yao, Z., Ai, J., Li, B., Zhang, C.: Efficient detr: improving end-to-end object detector with dense prior. arXiv preprint arXiv:2104.01318 (2021). https://doi.org/10.48550/arXiv.2104.01318
45. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6022–6031 (2019). https://doi.org/10.1109/ICCV.2019.00612
46. Zhang, H., Cisse, M., Dauphin, Y., Lopez-Paz, D.: mixup: Beyond empirical risk management. In: Proceedings of the International Conference on Learning Representations (ICLR). pp. 1–13 (2018). https://doi.org/10.48550/arXiv.1710.09412
47. Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., Liu, Z.: Weakly aligned cross-modal learning for multispectral pedestrian detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5126–5136 (2019). https://doi.org/10.1109/ICCV.2019.00523
48. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1259–1267 (2016). https://doi.org/10.1109/CVPR.2016.141
49. Zhou, C., Yuan, J.: Bi-box regression for pedestrian detection and occlusion estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 138–154 (2018). https://doi.org/10.1007/978-3-030-01246-5_9
50. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021). https://doi.org/10.48550/arXiv.2104.02008
51. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2242–2251 (2017). https://doi.org/10.1109/ICCV.2017.244
52. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020). https://doi.org/10.48550/arXiv.2010.04159