PIT: Position-Invariant Transform for Cross-FoV Domain Adaptation

Qiqi Gu[1*] Qianyu Zhou[1*] Minghao Xu[1] Zhengyang Feng[1]
Guangliang Cheng[2 5] Xuequan Lu[3 †] Jianping Shi[2 7] Lizhuang Ma[1 4 6 †]
[1]Shanghai Jiao Tong University, [2]SenseTime Group Research
[3]Deakin University, [4]East China Normal University, [5]Shanghai AI Laboratory
[6]MoE Key Lab of Artificial Intelligence, SJTU, [7]Qing Yuan Research Institute, SJTU
{miemie, zhouqianyu, xuminghao118, zyfeng97}@sjtu.edu.cn, guangliangcheng2014@gmail.com
xuequan.lu@deakin.edu.au, shijianping@sensetime.com, ma-lz@cs.sjtu.edu.cn

Abstract

Cross-domain object detection and semantic segmentation have witnessed impressive progress recently. Existing approaches mainly consider the domain shift resulting from external environments including the changes of background, illumination or weather, while distinct camera intrinsic parameters appear commonly in different domains and their influence for domain adaptation has been very rarely explored. In this paper, we observe that the Field of View (FoV) gap induces noticeable instance appearance differences between the source and target domains. We further discover that the FoV gap between two domains impairs domain adaptation performance under both the FoV-increasing (source FoV $<$ target FoV) and FoV-decreasing cases. Motivated by the observations, we propose the ***Position-Invariant Transform*** *(PIT) to better align images in different domains. We also introduce a reverse PIT for mapping the transformed/aligned images back to the original image space, and design a loss re-weighting strategy to accelerate the training process. Our method can be easily plugged into existing cross-domain detection/segmentation frameworks, while bringing about negligible computational overhead. Extensive experiments demonstrate that our method can soundly boost the performance on both cross-domain object detection and segmentation for state-of-the-art techniques. Our code is available at* https://github.com/sheepooo/PIT-Position-Invariant-Transform.

1. Introduction

Object detection [20, 41, 40] and semantic segmentation [33, 5, 6, 14] are two fundamental problems in computer vision. The former aims at precisely locating and identifying the objects in an image and the latter targets to classify the semantics of each pixel. Training a generalized model with high performance for the two tasks calls for massive images with elaborate annotations, while it is laborious to prepare such well-annotated data. Meanwhile, due to the existence of domain shift [1], a model trained on a specific dataset often suffers from significant performance degradation when applied to another domain. A common

(a) Original Images (b) Transformed Images

Figure 1: Objects (cars) in different positions relative to the camera have different extent of deformation, which remarkably burdens the alignment of intra-class features. This can be effectively mitigated by our PIT. Top row: images of an object (in different positions) captured by a virtual camera. Other rows: real photos from the KITTI dataset.

*Equal Contribution. † Joint Corresponding author.

solution is to transfer the knowledge acquired from a labeled source domain to an unlabeled target domain, which is known as Unsupervised Domain Adaptation (UDA) [38].

In general, two typical manners have been explored to adapt models from the source to the target domain. One is pixel-level alignment, target-like images are generated to provide implicit or explicit supervisory signals on target domain [44, 25, 61]. The other is feature-level alignment, the feature distributions of two domains are aligned through constraining domain discrepancy metrics [34, 48, 62] or performing feature confusion [18, 52, 39].

In the study of cross-domain detection/segmentation, previous works [8, 43, 60, 2, 9, 59, 25, 50, 31, 65] mainly focus on narrowing the domain shift caused by external environments, *e.g.* the change of background, illumination and weather, etc. However, very little attention has been paid to the camera's intrinsic parameters which often bring noticeable domain discrepancy due to the use of various cameras.

We observe that one main camera parameter, the Field of View (FoV)[1], induces a distinct dimension of the domain gap. As a matter of fact, the FoV discrepancy frequently occurs among datasets or in real-world scenarios. For instance, in autonomous driving, cameras with different FoVs are often used together, because of the inevitable updating of cameras in the long period of data collection. FoV difference derives the variety of instance structural appearances across the source and target domains, leading to the sample diversifying within a category. This obviously increases the burden of domain adaptation models, thus resulting in less desired performance.

Motivated by the above observation, in this paper we attempt to alleviate the adverse impact of the diverse FoVs between domains, in order to boost the performance of cross-domain detection/segmentation.

We discuss the influence of the FoV gap in two general cases. (1) In FoV-increasing adaptation (the FoV of the target domain is larger than that of the source domain), the target domain instances with large incident angles cannot be well aligned to the source domain for the lack of similar-appearance counterparts. (2) In FoV-decreasing adaptation (target FoV smaller than source FoV), the sparsity of the source domain instances within a specific range of incident angle also hampers domain alignment. Existing UDA methods usually try to bridge the whole domain gap and optimize the model without specifically taking account of the FoV factor, thus preventing the model from fully learning domain-invariant features.

To address the above problem, we propose the **Position-Invariant Transform** (PIT) to straightforwardly narrow the FoV gap between the source and target domains (Fig. 1). Specifically, the pixels lying in the original imaging space are mapped to another two-dimensional space shaped as a spherical surface, such that the appearances of the instances in various positions are aligned to a great extent. Also, we introduce a reverse PIT for mapping the transformed images back to the original image space. In addition, we design an efficient loss re-weighting strategy to speed up the training procedure. Our modules induce little computational overhead while boosting performance, and they can be easily served as plug-and-play modules to any existing cross-domain detection/segmentation frameworks.

Our contributions can be summarized as follows:

- We statistically analyze the negative influence of FoV difference between the source and target domains on UDA models, in which both the increasing and decreasing of FoV between domains impair the domain alignment.
- We propose the Position-Invariant Transform (PIT) to align instance structural appearances in different positions in each category, and reverse PIT to map the transformed images to the original image space. We also introduce a loss re-weighting strategy to speed up the training procedure.
- The effectiveness of PIT is verified on both cross-domain detection and segmentation tasks. Equipped with our modules, state-of-the-art UDA methods show soundly better performance than before.

2. Related Work

Unsupervised Domain Adaptation (UDA). UDA aims to adapt the model trained on a labeled source domain to an unlabeled target domain by reducing the distribution gap between two domains. A group of recent approaches focused on minimizing the domain discrepancy [34, 48, 62] metric (*e.g.* Maximum Mean Discrepancy [53]), adversarial learning [18, 52, 39] or prototype-based alignment [58, 42, 60]. Despite the successes achieved in classification-based tasks [34, 18, 48, 52, 44, 61], these methods work well on simple classification datasets (*e.g.* MNIST [30] and SVHN [37]), but can hardly be applied to more challenging tasks, *e.g.* object detection and semantic segmentation.

Domain Adaptive Detection/Segmentation. Not until recently has the community paid attention to domain shift problem in object detection or semantic segmentation. This line of research has been investigated by a large number of researchers, and great efforts have been made to explore a variety of algorithms and architectures to reduce the domain gap in pixel-level [25, 3, 31, 29, 21], feature-level [35, 69, 2, 8, 66], instance-level [8, 59, 4] and output-level [50, 51, 36, 63], which have shown successes on both object detection [8, 67, 2, 43, 4, 60, 59]

[1]Field of View (FoV): in photography, the angle between two rays passing through the perspective center (rear nodal point) of a camera lens to the two opposite sides of the format.

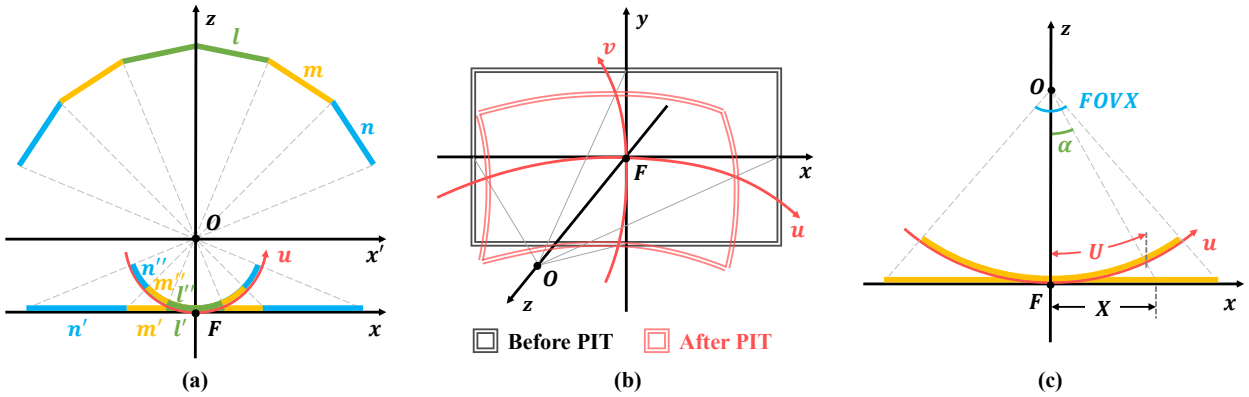

Figure 2: **(a)** Illustration of the position-related deformation and the position-invariance of PIT. **(b)** The 3D spatial relationship of images before and after PIT. **(c)** The transformation between two coordinate systems. O: the optical center of a camera; F: the focal point; $x'Oy'$: the plane of lens (y' axis is perpendicular to $x'Oz$); xFy: the imaging plane which is parallel to $x'Oy'$; uFv: the spherical surface to map the image, where the coordinate axes u and v are arcs.

and semantic segmentation [25, 50, 36, 31, 57, 51, 55, 65]. The current mainstream approaches of these two tasks include adversarial learning [67, 23, 43, 50, 36, 55, 51], self-training [69, 68, 28] and self-ensembling [2, 9, 13, 64, 65]. Despite the great progress, these works mainly focused on adapting different external environmental conditions, *e.g.* background, illumination and weather. While the gap of camera intrinsic parameters between distinct domains has been ignored. In this work, we show the effectiveness of our method by easily integrating it into adversarial learning and self-ensembling on these two tasks.

CNNs with Geometric Transformations. Researchers investigated CNNs with the abilities of geometric transformation or deformation gains over the past years. Spatial transformer networks [26] predicted the transformation parameters to reduce the influence of affine transformations. Active convolution [27] designed a transformable convolution kernel to get a more general shape of receptive field. Deformable convolution network [11] further improved the former by predicting the receptive field location, and [47] used spherical CNN to translate a planar CNN to process 360° imagery directly in its equirectangular projection. Largely different from these methods which mainly focused on designing new network architectures, our method pays more attention to the attribution of the data itself (*i.e.* position-related deformation caused by camera imaging) to enhance the feature alignment in UDA models.

3. Method

In Unsupervised Domain Adaptation (UDA), a source domain $\mathcal{S} = \{(x_i^{\mathcal{S}}, y_i^{\mathcal{S}})\}_{i=1}^{N_{\mathcal{S}}}$ with $N_{\mathcal{S}}$ labeled samples and a target domain $\mathcal{T} = \{x_j^{\mathcal{T}}\}_{j=1}^{N_{\mathcal{T}}}$ with $N_{\mathcal{T}}$ unlabeled samples are available, where $x_i^{\mathcal{S}}$ follows source distribution $\mathbb{P}_{\mathcal{S}}$, and $x_j^{\mathcal{T}}$ obeys target distribution $\mathbb{P}_{\mathcal{T}}$. The objective of UDA is to train a model generalizing well in the target domain, using the above data from both domains.

3.1. Motivation

In the real world, images are often captured by cameras with distinct intrinsic parameters, which leads to the cross-camera domain gap. We observe that the structure of objects deform noticeably as their positions change, and the FoV parameter mainly impacts the deformation extent (Fig. 1).

The FoV parameter restricts the angle of the area that can be observed by a camera, *i.e.* the maximum incident angle of observable objects. Fig. 2 (a) illustrates how the variance of the incident angle affects the structural appearance of an object. l, m, and n are structure-alike objects which lie in different positions with the same distance to the optical center O. When projected onto the imaging plane, the length of their images l', m', and n' are obviously different. Specifically, with the increase of an object's deviation from the center of a scene (*i.e.* the expansion of the incident angle), its camera imaging becomes longer, which makes the object structure vary in different positions of an image.

Because of the restriction on the range of incident angles by FoV, the structural appearance of objects within the same category can be noticeably distinct between the source and target domains, as shown in Fig. 1 where different degrees of imaging deformation may occur in two domains. This kind of deformation is totally different from the lens distortion [54]. The latter is a deviation from rectilinear projection and can be fixed by camera calibration, and the calibrated image is the ideal projection on the imaging plane.

Considering the significance of learning structure-invariant feature representations in scene understanding [26, 11], the structural difference between the objects from two domains can trap a UDA model into a dilemma in which

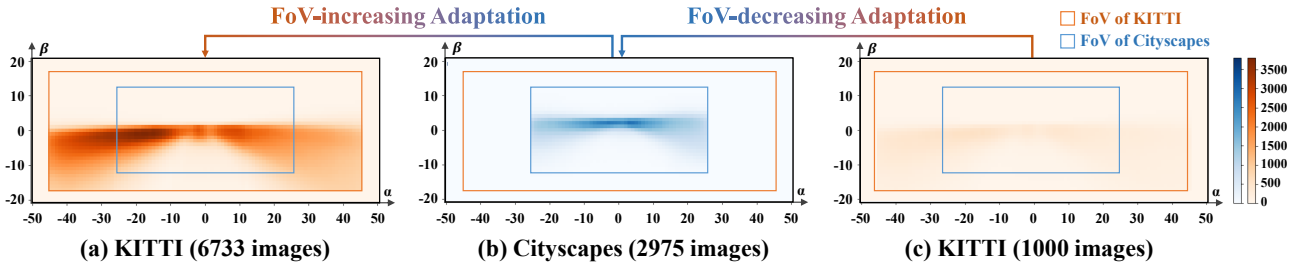

Figure 3: Heat maps of foreground occurrences for each (α, β) in AP space. **(a)(c)** Statistics from the whole KITTI [19] training set and a subset involving 1,000 images, which stand for large-scale and small-scale datasets in the real world, respectively. **(b)** Statistics from the Cityscapes training set.

that kind of difference cannot be handled well. To better elucidate the existence and underlying effect of the FoV gap between the two domains, we statistically analyze the incident angle distribution in various datasets. Specifically, we define α and β (α is shown in Fig. 2 (c), β is the counterpart in yFz plane) as a point's incident angles towards the optical center along the horizontal and vertical axis, respectively. Notice that the imaging deformation of an object is closely related to α and β, and the deformation extent continuously increases along these two angles' absolute values. Therefore, we span the α axis and β axis to form a two-dimensional space, named as **Angular-Position space** (AP space), in which the absolute value of each point's coordinate measures the horizontal and vertical deformation extent of the object lying in the corresponding position. We then count the number of foreground occurrences for each (α, β) integer values on KITTI [19] and Cityscapes [10] datasets, and these statistics are displayed as heat maps in Fig. 3. It can be observed that the objects of KITTI dataset distribute in a wider range of incident angles than those of Cityscapes dataset, which leads to two opposite directions of cross-FoV adaptation (see below).

FoV-increasing Adaptation. In this case, the target domain possesses a wider FoV distribution than the source domain, *e.g.* adapting from Cityscapes (Fig. 3(b)) to KITTI (Fig. 3(a)), which means that the objects in target domain own a greater range of deformation extents. Consequently, some target objects fall in the regions without source objects in the AP space, and they cannot be well aligned to the source domain for the lack of proper supervision from similar-appearance counterparts, which impairs the performance of UDA models. The proposed PIT module (Sec. 3.2) effectively mitigates this defect via its position-invariance.

FoV-decreasing Adaptation. In this case, the target domain has a narrower FoV distribution, *e.g.* adapting from KITTI (Fig. 3(a)(c)) to Cityscapes (Fig. 3(b)), such that the distributional range of target objects are covered by that of source objects. It is true that when the source objects are dense enough everywhere (Fig. 3(a)) in the AP space, domain alignment can be well performed by a UDA method. However, when the source domain has low data density (Fig. 3(c)), a target object can hardly find its source counterparts with a similar structural appearance which it can align with; meanwhile the source samples are not fully utilized. Under this situation, the proposed PIT approach (Sec 3.2) is able to gather source objects in the AP space and thus eases the alignment.

3.2. Position-Invariant Transform

The object deviating more from the principal axis of the lens would be stretched to a greater extent in the camera imaging process, which manifests the imaging deformation phenomenon in Fig. 1.

In order to alleviate this kind of deformation, we propose the **Position-Invariant Transform** (PIT). Fig. 2(a) shows the principle of PIT. The location of a point's image is the intersection of its incident light passing through the optical center O and the imaging surface, so the imaging of a scene would be altered by changing the imaging surface. In this method, the incident light from an object passing through O is received with a spherical surface instead of a plane, i.e. the uFv surface with sphere center O shown in Fig. 2(b). In such a spherical space, images can largely retain the relative size of original objects. For the same-size objects l, m and n in Fig. 2(a), they are mapped to l'', m'' and n'' with the same length on the uFv surface. This example illustrates that the imaging on a spherical surface is invariant to the object's angular position, *i.e.* satisfying *position-invariance*.

After manifesting the benefits of a spherical surface over 2D imaging plane, a projection from spherical image back to a new plane image is needed to match the image with the input form of network. Thus, such projection should have two properties, which cannot be satisfied by the existing projection approaches (e.g. equirectangular, Mercator, etc.): (1) the image space after transformation should obey position-invariance, in order to align instances in the pixel level; (2) the horizontal (vertical) line should remain horizontal (vertical) after transformation, so as to ensure the

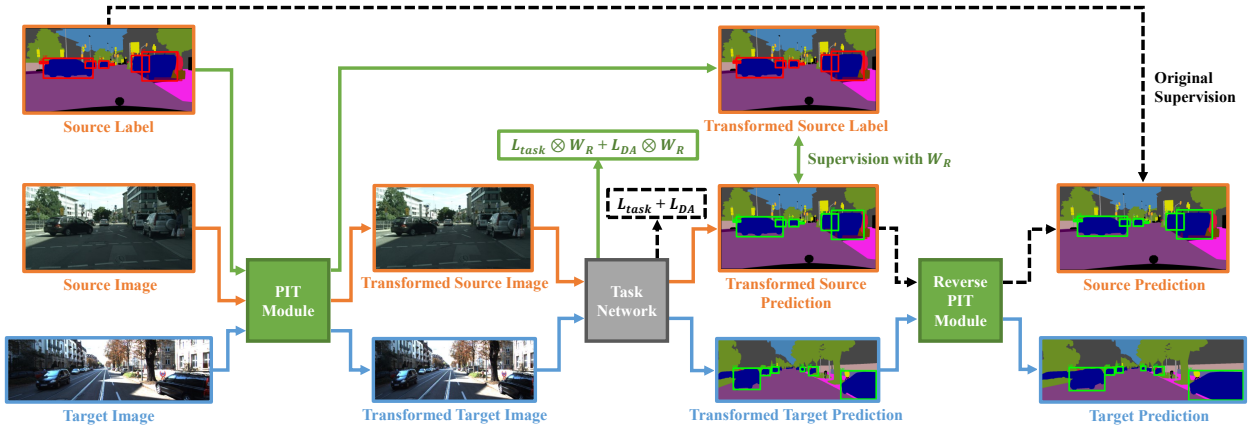

Figure 4: Overview of our method.

validity of bounding box labels. Taking both properties into consideration, we formulate a new projection which is defined as (referring to Fig. 2(c) for intuitive notions):

$$X(U) = f \times \tan(\frac{U}{f}), \tag{1}$$

$$Y(V) = f \times \tan(\frac{V}{f}), \tag{2}$$

$$M'[U][V] = M[X(U)][Y(V)], \tag{3}$$

where (X, Y) is the coordinate in the original image space (*i.e.* the xFy coordinate system with origin F), and (U, V) is the coordinate in the image space after PIT (*i.e.* the uFv coordinate system with origin F). $M[X][Y]$ and $M'[U][V]$ denote the pixel values of the corresponding points before and after transformation. f is the focal length which can be estimated using the FoV parameter or precisely calculated by camera calibration.

As shown in Fig. 1, the size of an image becomes smaller after PIT, and the regions further from the center of a scene are compressed with a higher ratio. Furthermore, the vertical/horizontal lines are preserved after PIT.

3.3. Cross-FoV Domain Adaptation

Integration. The proposed PIT method can be utilized as a plug-and-play module to existing cross-domain detection and segmentation frameworks. As shown in Fig. 4, both the images from the source and target domains are first fed into the PIT module to be transformed into the position-invariant ones, which serve as the inputs to the task network. In the training phase, the labels from source domain are also transformed by PIT to provide supervision. As for inference, the prediction result of the task network is mapped back to the original image space by the reverse PIT module which outputs the final prediction.

Reverse PIT and loss re-weighting strategy. Since the evaluation is conducted with the un-transformed ground truth, it is plausible to provide supervision with the original labels, as shown by the black dash lines in Fig. 4. However, different from the PIT process which only needs to execute once for each input image in the datasets, the reverse PIT module would be employed in each iteration and cause extra computational cost. In order to accelerate the training, we design a pixel-wise loss re-weighting strategy to substitute the reverse PIT module during the training process. A pixel in the transformed image corresponds to a region in the original image, and each pixel in the original image weighs equally in evaluation. Therefore, a transformed pixel's weight should be the area of its mapping region, depending on the pixel's position. With this weight, the transformed supervision is equivalent to the reverse PIT in terms of loss computation. The weighting matrix is formulated as:

$$w_R(U, V) = (X(|U|+1) - X(|U|)) \times (Y(|V|+1) - Y(|V|)), \tag{4}$$

where w_R is the weight assigned to pixel located in (U, V) in the transformed image.

Using the weights derived above, we re-weigh the pixel-wise losses, including the task-specific loss L_{task} (*e.g.* the supervised loss L_{sup} in [9]) and the domain adaptation loss L_{da} (*e.g.* the consistency loss L_{con} in [9]):

$$L = L_{task} \otimes W_R + \lambda L_{da} \otimes W_R, \tag{5}$$

where λ is the weight to balance the two losses.

With this loss re-weighting strategy, we can use the transformed labels to optimize the model, as shown by the green line in Fig. 4, which speeds up the training procedure.

4. Experiments

We conduct extensive experiments on object detection and semantic segmentation tasks. The results show that our

approach can soundly boost the performance on cross-FoV adaptation by easily plugging it into any UDA frameworks.

4.1. Experimental Setup

Datasets. We utilize three public datasets provided with FoV parameters in our experiments: Cityscapes [10], KITTI [19] and Virtual KITTI [17]. In here, we add a two-dimensional array after the name of each dataset, to indicate the approximate horizontal and vertical FoV parameters $(FoVx, FoVy)$ of the camera for scene capturing.

- **Cityscapes** [10] $(50°, 26°)$ is a dataset of street scenes in several cities. It owns 2,975 images for training and 500 for validation, and both of them have dense pixel-level labels. We get the bounding box labels for object detection task by calculating the tightest rectangles of instance annotations as [8] did. It uses 4 types of cameras with different FoVs ($49.5° < FoVx < 51.7°$, $25.5° < FoVy < 26.2°$), and we process each image with its own recorded FoV.
- **KITTI** [19] $(90°, 34°)$ is a real-world dataset containing 7,481 images with bounding boxes and another 200 images with pixel-level labels. In the detection task, we split the the training set and the validation set manually. In the segmentation task, it is used as the target domain only due to the lack of pixel-level annotations.
- **Virtual KITTI** [17] $(80°, 29°)$ is a synthetic dataset which clones the scenes from the KITTI with 21,260 images. It provides pixel-level instance labels, and the bounding boxes are obtained as those in Cityscapes.

Baselines and Comparison Methods. Following the experimental design in [59], we select SWDA [43], SCL [45], GPA [60] as our baseline methods for cross-domain detection, and Self-Ensembling [9], CowMix[16], CutMix[15], DACS [49] for cross-domain segmentation.

We re-implement these methods for fair comparisons, and our re-implementations attain higher accuracies than the reported ones. When comparing with other state-of-the-art methods, we use the results from the original papers.

Implementation Details. In object detection experiments, VGG16 [46] model pre-trained on ImageNet [12] is used as the backbone of all the selected methods. The hyper-parameters are set according to the original papers. The average precision (AP) is used as evaluation metric.

In semantic segmentation experiments, the DeepLab-v2 [5] with ResNet101 [22] pretrained on ImageNet [12] and on MSCOCO [32] is used as our backbone. Hyper-parameters are set following [50, 9].

4.2. The Existence of FoV Gap

In order to prove the existence of FoV gap, we crop the images (Fig. 5) to generate new datasets with certain $FoVx$. Then we train a Faster-RCNN [41] model (but **NOT** a UDA method) on KITTI-50°, and test it directly on KITTI-70°/80°/90°to examine the compactness of features.

Table 1: Source-only detection results (car AP, %) traind on KITTI-50°and tested on different degrees ($FoVx$) of cropped KITTI.

FoVx	50°	70°	80°	90°
FR [41]	87.49%	86.80%	86.31%	84.92%
FR + PIT	**87.81%**	**87.37%**	**86.88%**	**86.43%**

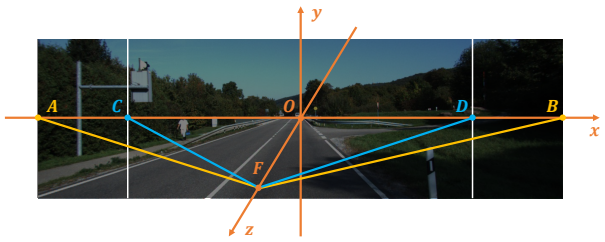

Figure 5: Cropping image with certain $FoVx$. $FoVx$ was reduced from $\angle AFB$ to $\angle CFD$ after cropping.

Tab. 1 shows the detection results of these source-only experiments. Without PIT, the performance gets worse as the FoV gap gets bigger, while PIT effectively suppresses the performance drop. It demonstrates that the PIT module plays an important role in bridging the FoV gap.

4.3. Domain Adaptation for Object Detection

4.3.1 FoV-increasing Adaptation

Cityscapes $(50°, 26°) \rightarrow$ **KITTI** $(90°, 34°)$**.** It's a cross-camera adaptation, in which FoV gap is one of the main components of the domain gap. Table 3 shows the AP results of the car class. With our proposed PIT method, all the methods performed much better than their vanilla versions. The highest gain reaches 5.27%, which is a remarkable improvement in object detection.

Virtual KITTI $(80°, 29°) \rightarrow$ **KITTI** $(90°, 34°)$**.** It's a synthetic-to-real adaptation in which FoV gap is a minor factor of domain gap. The results are shown in Table 4.

In order to look into the factors which influence the effectiveness of our method, we design controlled experiments. We crop the images (see Fig. 5) with certain $FoVx$ and use them as the source or target domain.

In Table 2, experiments in the upper part have the same source $FoVx$ and incremental target $FoVx$ (*i.e.*, incremental FoV gap), and those in the bottom part of Table 2 have a constant target $FoVx$ with different source $FoVx$. With the fixed $FoVx$ in one domain, the larger $FoVx$ gaps result in worse performance in the baseline, while our method gains higher improvement. These results verify that our pro-

Table 2: Detection results of Virtual KITTI → KITTI (cropped with specified $FoVx$.)

Source $FoVx$	Method	Target $FoVx$: 50°		Target $FoVx$: 70°		Target $FoVx$: 90°	
		car AP(%)	Gain(%)	car AP(%)	Gain(%)	car AP(%)	Gain(%)
40°	SCL (Arxiv'20) [45]	69.81	**0.75**	68.04	**1.04**	65.71	**3.33**
	SCL + PIT	**70.56**		**69.08**		**69.04**	
Target $FoVx$	Method	Source $FoVx$: 40°		Source $FoVx$: 60°		Source $FoVx$: 80°	
		car AP(%)	Gain(%)	car AP(%)	Gain(%)	car AP(%)	Gain(%)
90°	SCL (Arxiv'20) [45]	65.71	**3.33**	67.74	**2.46**	70.50	**1.41**
	SCL + PIT	**69.04**		**70.20**		**71.91**	

Table 3: Detection results of Cityscapes → KITTI.

Methods	car AP(%)	Gain(%)
DAFRCN*(CVPR'18) [8]	64.10	-
SWDA**(CVPR'19) [43]	71.00	-
MAF*(ICCV'19) [23]	72.10	-
SCL*(Arxiv'19) [45]	72.70	-
ATF*(ECCV'20) [24]	73.50	-
SWDA (CVPR'19) [43]	72.42	**3.35**
SWDA + PIT	**75.77**	
SCL (Arxiv'19) [45]	75.28	**1.84**
SCL + PIT	**77.11**	
GPA (CVPR'20) [60]	69.24	**5.27**
GPA + PIT	**74.51**	

* reported from its original paper, and ** from [24].

Table 4: Detection results of Virtual KITTI → KITTI.

Methods	car AP(%)	Gain(%)
SWDA (CVPR'19) [43]	69.74	**2.12**
SWDA + PIT	**71.86**	
SCL (Arxiv'19) [45]	70.50	**1.41**
SCL + PIT	**71.91**	
GPA (CVPR'20) [60]	65.36	**5.35**
GPA + PIT	**70.71**	

posed method can effectively narrow the specific FoV gap.

4.3.2 FoV-decreasing Adaptation

As analyzed in Section 3.1, in this case, our method works with insufficient labeled data. So we reduce the size of the source dataset manually for the experiment setting, with no special treatment on the target domain.

KITTI $(90°, 34°)$ → **Cityscapes** $(50°, 26°)$. We use 1,000 labeled images in KITTI dataset as source data. Table 6a shows the detection results on Cityscapes, and our method outperforms baselines by $1.48\% \sim 2.01\%$ on car AP.

Virtual KITTI $(80°, 29°)$ → **Cityscapes** $(50°, 26°)$. We use the "clone" subset (2126 images) of Virtual KITTI as source data. As shown in Table 6b, our method achieves increases when plugged in all the baseline networks.

Table 5: Comparison with data augmentation.

SWDA [43]	Aug	PIT	car AP(%)	Gain(%)
✓			72.42	-
✓	✓		74.74	2.32
✓		✓	75.77	3.35
✓	✓	✓	76.93	4.51

4.4. Domain Adaptation for Semantic Segmentation

We conduct two experiments : 1) Cityscapes $(50°, 26°)$ → KITTI $(90°, 34°)$, 2) Virtual KITTI $(80°, 29°)$ → KITTI$(90°, 34°)$. mIoUs are reported for comparisons. The class-wise IoUs are reported in the supplementary material.

The results are shown in Table 7. Assembled in four state-of-the-art domain adaptative semantic segmentation methods, our method improves the mIoUs by 1.06% to 1.77% compared to the original methods, which again demonstrates the effectiveness of the proposed method.

4.5. Comparison with Data Augmentation

Though served as a fixed part before and after the network, PIT is totally different from data augmentation. Data augmentation processes data with random parameters in several directions to diversify samples, while PIT aims at the opposite purpose. It calculates the optimal transform directly and reduces the variety of intra-class instances, which is beneficial for the feature alignment.

We use the commonly-used data augmentation (random scale and random crop) [56] in experiments. Tab. 5 shows

Table 6: Detection results of FoV-decreasing case.

(a) KITTI subset → Cityscapes.

Methods	car AP(%)	Gain(%)
SWDA (CVPR'19) [43]	39.67	**2.01**
SWDA + PIT	**41.68**	
SCL (Arxiv'20) [45]	38.64	**1.61**
SCL + PIT	**40.25**	
GPA (CVPR'20) [60]	44.77	**1.48**
GPA + PIT	**46.25**	

(b) Virtual KITTI subset → Cityscapes.

Methods	car AP(%)	Gain(%)
SWDA (CVPR'19) [43]	37.53	**1.42**
SWDA + PIT	**38.95**	
SCL (Arxiv'19) [45]	37.22	**1.49**
SCL + PIT	**38.71**	
GPA (CVPR'20) [60]	44.56	**1.00**
GPA + PIT	**45.56**	

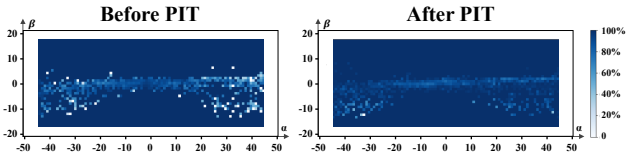

Figure 6: The bin-wise performance in AP space.Based on detection task Cityscapes → KITTI, SWDA [43] backbone.

Table 7: Segmentation results.

(a) Cityscapes → KITTI.

Method	mIoU	Gain
Self-Ensembling (ICCV'19) [9]	59.54	**1.45**
Self-Ensembling + PIT	**61.00**	
CowMix (Arxiv'20) [16]	59.15	**1.22**
CowMix + PIT	**60.37**	
CutMix (BMVC'20) [15]	58.78	**1.31**
CutMix + PIT	**60.09**	
DACS (WACV'21) [49]	59.19	**1.63**
DACS + PIT	**60.82**	

(b) Virtual KITTI → KITTI.

Method	mIoU	Gain
GIO-Ada* (CVPR'19) [7]	53.5	-
Self-Ensembling (ICCV'19) [9]	55.45	**1.77**
Self-Ensembling + PIT	**57.22**	
CowMix (Arxiv'20) [16]	56.07	**1.17**
CowMix + PIT	**57.24**	
CutMix (BMVC'20) [15]	55.58	**1.14**
CutMix + PIT	**56.72**	
DACS (WACV'21) [49]	55.51	**1.06**
DACS + PIT	**56.57**	

that PIT and data augmentation play different roles in the UDA task. Data augmentation aims at the linear transformation of objects (*e.g.* different scale), while PIT reduces the instance diversity caused by non-linear deformations.

4.6. Visualization

In order to demonstrate the effectiveness of PIT over different incident angles, we reported the performance for different bins in the AP space (specified in Sec. 3.1). Using the center point to represent a predicted bounding box, we calculate the bin-wise accuracy and visualize in Fig. 6. There are clear improvements in the peripheral regions where the objects have greater deformation, which verifies the effectiveness of the instance alignment through PIT. See more results in the supplementary material.

5. Conclusion

In this paper, we statistically analyzed the impact of FoV difference between domains, including both FoV-increasing and -decreasing cases. Then we proposed a novel method (PIT) for cross-FoV detection/segmentation, which can be widely used in real-world applications due to the variety of cameras. Our method aligns the structural appearance of instances in the same category across domains. We also design a loss re-weighting strategy as a substitution of reverse PIT to speed up the training. As a plug-and-play approach, our method can be easily embedded into a wide range of existing networks. Experiments demonstrate that it boosts the performance in cross-domain detection and segmentation.

6. Acknowledgement

This work is supported by National Key Research and Development Program of China (No. 2019YFC1521104), Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0102), Shanghai Science and Technology Commission (No. 21511101200), Zhejiang Lab (No. 2020NB0AB01) and National Natural Science Foundation of China (No. 61972157 and No. 62106268). The author Qianyu Zhou is supported by Wu Wenjun Honorary Doctoral Scholarship, AI Institute, Shanghai Jiao Tong University.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

[2] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*, 2019.

[3] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *CVPR*, 2019.

[4] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, 2020.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[7] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, 2019.

[8] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018.

[9] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *ICCV*, 2019.

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[13] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross domain object detection. *arXiv preprint arXiv:2003.00707*, 2020.

[14] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and classbalanced curriculum. *arXiv preprint arXiv:2004.08514*, 2020.

[15] Geoff French, Samuli Laine, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*, 2020.

[16] Geoff French, Avital Oliver, and Tim Salimans. Milking cowmask for semi-supervised image classification. *arXiv preprint arXiv:2003.12022*, 2020.

[17] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016.

[18] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

[19] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJR*, 32(11):1231–1237, 2013.

[20] Ross Girshick. Fast r-cnn. In *CVPR*, 2015.

[21] Shaohua Guo, Qianyu Zhou, Ye Zhou, Qiqi Gu, Junshu Tang, Zhengyang Feng, and Lizhuang Ma. Label-free regional consistency for image-to-image translation. In *ICME*, 2021.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[23] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, 2019.

[24] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. *arXiv preprint arXiv:2007.01571*, 2020.

[25] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018.

[26] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, 2015.

[27] Yunho Jeon and Junmo Kim. Active convolution: Learning the shape of convolution for image classification. In *CVPR*, 2017.

[28] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *ICCV*, 2019.

[29] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *CVPR*, 2020.

[30] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[31] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 2019.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[34] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

[35] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *ICCV*, 2019.

[36] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019.

[37] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[38] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010.

[39] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *AAAI*, 2018.

[40] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[42] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, 2019.

[43] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, 2019.

[44] Swami Sankaranarayanan, Yogesh Balaji, Carlos D. Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *CVPR*, 2018.

[45] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[47] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360° imagery. In *NIPS*, 2017.

[48] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016.

[49] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. *arXiv preprint arXiv:2007.08702*, 2020.

[50] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018.

[51] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *ICCV*, 2019.

[52] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[53] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *Computer Science*, 2014.

[54] Paul van Walree. Distortion. photographic optics, 2009.

[55] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019.

[56] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q. Weinberger, and Wei-Lun Chao. Train in germany, test in the usa: Making 3d object detectors generalize. In *CVPR*, 2020.

[57] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *CVPR*, 2020.

[58] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *ICML*, 2018.

[59] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, 2020.

[60] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, 2020.

[61] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, 2020.

[62] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *CVPR*, 2017.

[63] Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. In *AAAI*, 2020.

[64] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. *arXiv preprint arXiv:2004.08878*, 2020.

[65] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Jiangmiao Pang, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Context-aware mixup for domain adaptive semantic segmentation. *arXiv preprint arXiv:2108.03557*, 2021.

[66] Qianyu Zhou, Qiqi Gu, Jiangmiao Pang, Zhengyang Feng, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Self-adversarial disentangling for specific domain adaptation. *arXiv preprint arXiv:2108.03553*, 2021.

[67] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, 2019.

[68] Yang Zou, Zhiding Yu, Xiaofeng Liu, B.V.K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, 2019.

[69] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018.