ISCAS 中国科学院软件研究所学术年会'2022
暨计算机科学国家重点实验室开放周

学术论文

自适应多粒度对齐的目标检测方法

Multi-Granularity Alignment Domain Adaptation for Object Detection

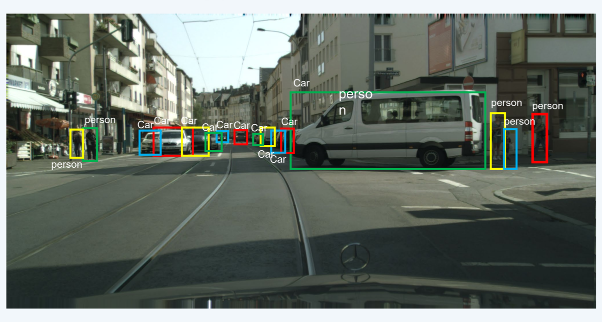
Wenzhang Zhou[1] Libo Zhang[1,2]*, Tiejian Luo[1], Yanjun Wu[2], [1]UCAS [2]ISCAS

CVPR 2022，张立波 18655882017

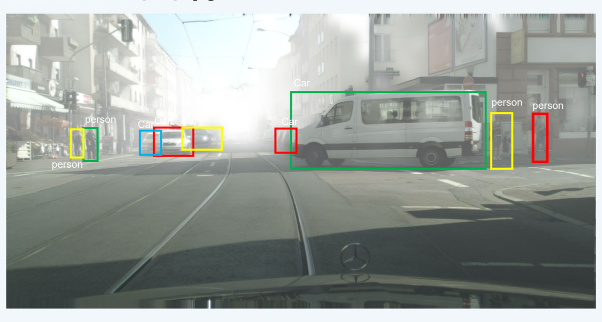
Introduction

- **Task:** Unsupervised adaptive detection is to improve the performance of detector learned from labeled source domain on new environments without labeled training data.
- **Solution**: the domain discriminator identifies whether the image is from source domain or target domain; while the object detector learns domain-invariant features to confuse the discriminator.

Cityscapes->Cityscapes
mAP: 34.7%



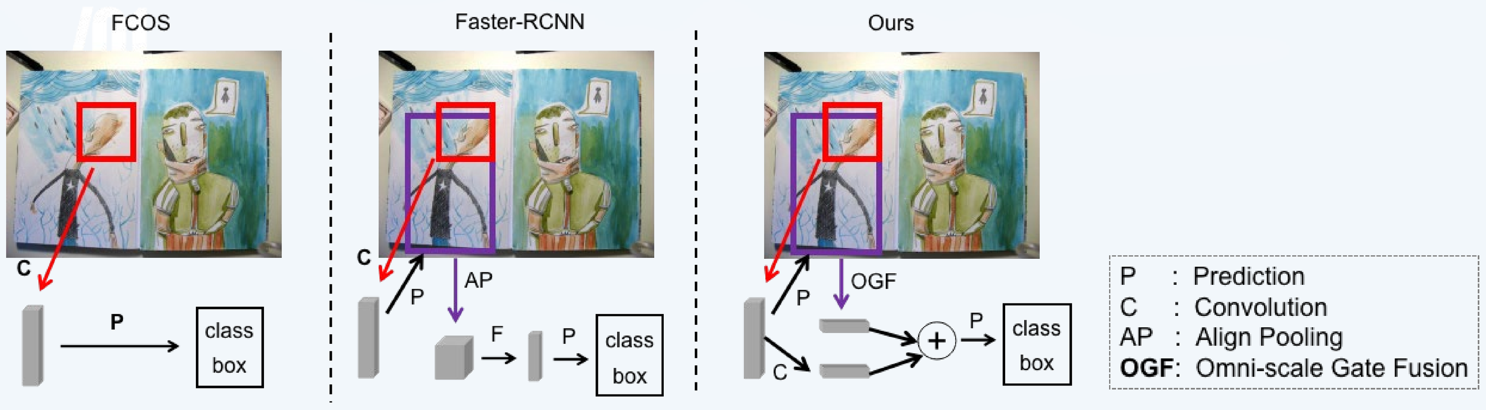
Cityscapes->Foggy Cityscapes
mAP: 18.8%



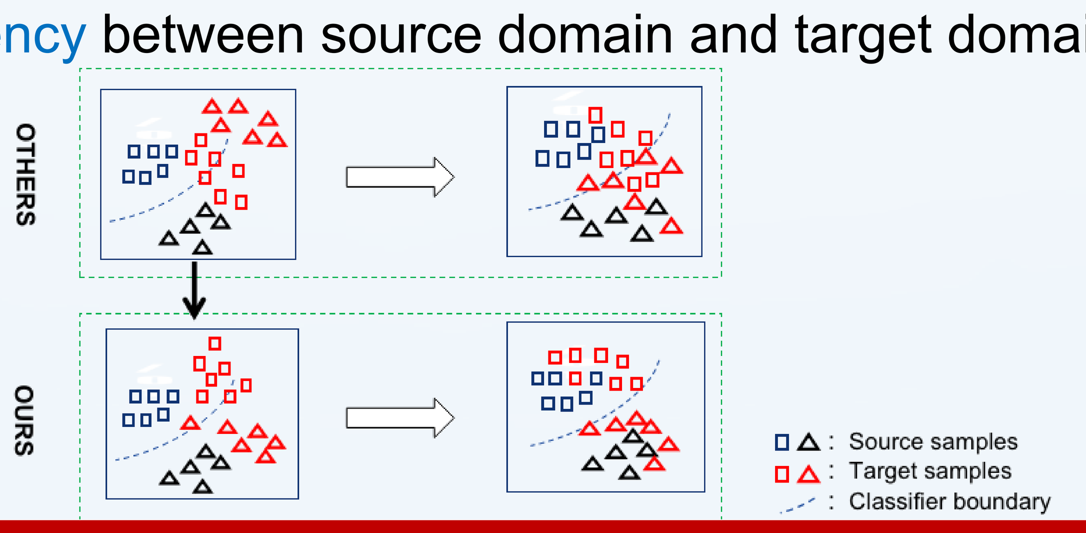
- **Challenges:** discrepancies in different scenes

Motivation

- **Discriminative representation:** The omni-scale gated fusion module can extract a discriminative representation in terms of objects with different scales and aspect ratios.

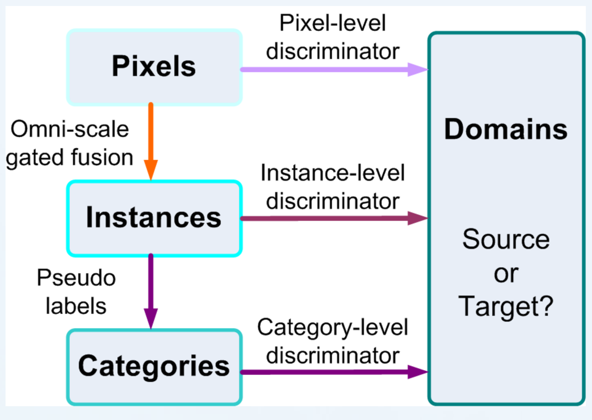


- **Distribution alignment:** The proposed category-level discriminator is to align the feature distribution based on instance discriminability in different categories and category consistency between source domain and target domain.

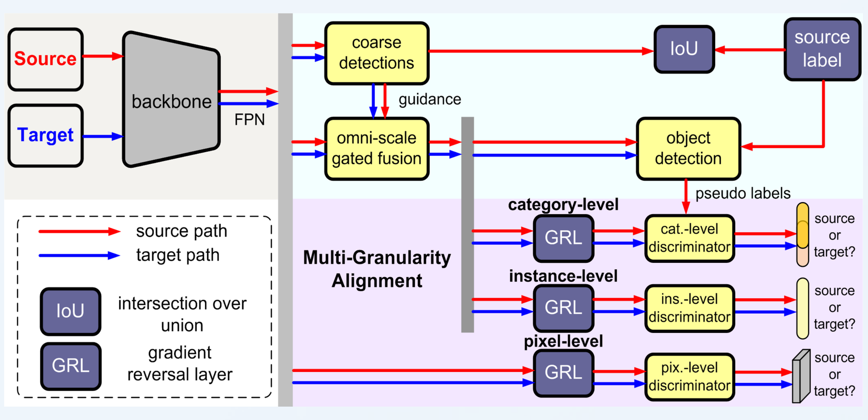


Our Approach

- **Our framework to encode multi-level dependencies**



- **Architecture of our domain adaptive object detection**

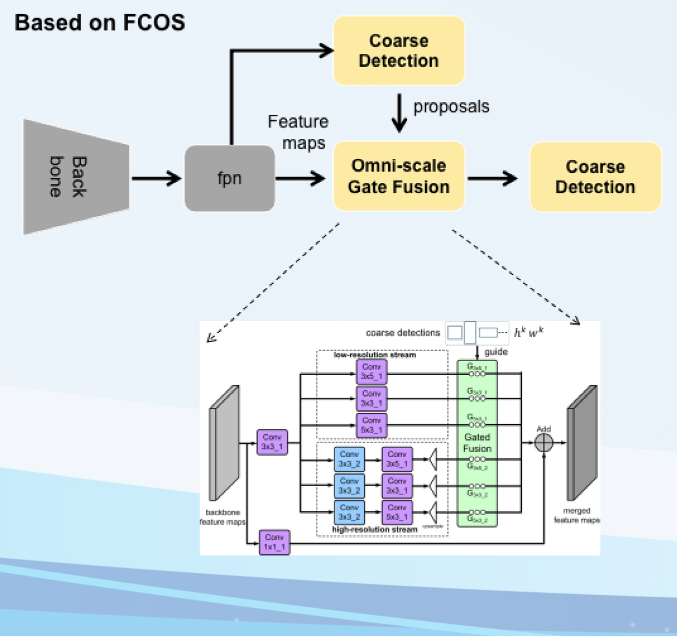


- **Omni-scale Gated Fusion**

For fcos framework, a series of convolutional layers and IoU is used as the coarse detectorand loss function, repsectively. To extend our framework to Faster-RCNN, we replace them with RPN and the original RPN loss.

$$\mathcal{L}_{\text{gui}} = -\sum_{k}\sum_{(i,j)} \ln(\text{IoU}(\tilde{b}^k_{i,j}, b^k_{i,j})) \text{ or } \mathcal{L}_{\text{gui}} = \mathcal{L}_{\text{rpn}}$$

where τ is the tempercature factor. o_ω denotesthe overlap between the predicted box and the convolution kernel ω. $\hat{o}$ is the maximal overlap among them.



- **Multi-Granularity Alignment**

— **Pixel-level and instance-level discriminators**

Pixel- and instance-level discriminator s are used to perform pixel and instan ce-level alignment of feature maps respectively. (Lpix and Lins employ the same loss function)

— **Category-level discriminator**

✓ **Instance Discriminability**

$$\mathcal{L}_{\text{dis}} = -\frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}\sum_{c=0}^{C-1} \hat{y}^{\text{dis}}_{i,j,c}\log(p^{\text{dis}}_{i,j,c})$$

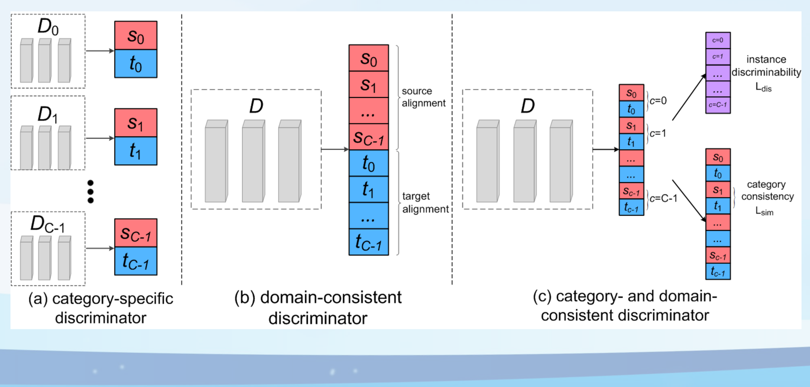
$$p^{\text{dis}}_{i,j,c} = \frac{\exp(\hat{M}_{i,j,2c} + \hat{M}_{i,j,2c+1})}{\sum_{c=0}^{C-1}\exp(\hat{M}_{i,j,2c} + \hat{M}_{i,j,2c+1})}$$

where $\hat{M}_{i,j,2c}$ and $\hat{M}_{i,j,2c+1}$ denote the confidence of the c-th category in source and target domains respectively

✓ **Category Consistency**

$$\mathcal{L}_{\text{sim}} = -\frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}\sum_{m=0}^{2C-1} \hat{y}^{\text{sim}}_{i,j,m}\log(p^{\text{sim}}_{i,j,m})$$

$$p^{\text{dis}}_{i,j,c} = \frac{\exp(\hat{M}_{i,j,2c} + \hat{M}_{i,j,2c+1})}{\sum_{c=0}^{C-1}\exp(\hat{M}_{i,j,2c} + \hat{M}_{i,j,2c+1})}$$



Experiments

- **Comparison with the state-of-the-art methods**

Cityscapes-> Foggy Cityscapes

method	detector	Backbone	persion	rider	car	truck	bus	train	mbike	bicycle	mAP
SAPNet	FRCNN	VGG-16	40.8	46.7	59.8	24.3	46.8	37.5	30.4	40.7	40.9
UMT	FRCNN	VGG-16	**56.5**	37.3	48.6	30.4	33.0	46.7	**46.8**	34.1	41.7
MeGA-CDA	FRCNN	VGG-16	37.7	49.0	52.4	25.4	49.2	**46.9**	34.5	39.0	41.8
CDG	FRCNN	VGG-16	38.0	47.4	53.1	**34.2**	47.5	41.1	38.3	38.9	42.3
ours	FRCNN	VGG-16	43.9	**49.6**	**60.6**	29.6	**50.7**	39.0	38.3	**42.8**	**44.3**
oracle	FRCNN	VGG-16	46.5	51.3	65.2	32.6	49.9	34.2	39.6	45.8	45.6
SST-AL	FCOS	-	45.1	47.4	59.4	24.5	50.0	25.7	26.0	38.7	39.6
CFA	FCOS	VGG-16	41.9	38.7	56.7	22.6	41.5	26.8	24.6	35.5	36.0
CFA	FCOS	ResNet-101	41.5	43.6	57.1	29.4	44.9	39.7	29.0	36.1	40.2
ours	FCOS	VGG-16	**45.7**	**47.5**	60.6	**31.0**	52.9	44.5	**29.0**	**38.0**	43.6
ours	FCOS	ResNet-101	43.1	47.3	**61.5**	30.2	**53.2**	**50.3**	27.9	36.9	**43.8**
oracle	FCOS	VGG-16	50.1	46.4	68.0	33.7	54.5	38.7	30.7	39.7	45.2
oracle	FCOS	ResNet-101	46.6	45.4	66.1	33.6	54.1	62.9	29.0	37.1	46.9

Sim10k/KITTI->Cityscapes

method	detector	Backbone	mAP
CST	FRCNN	VGG-16	44.5/43.6
MeGA-CDA	FRCNN	VGG-16	44.8/43.0
SAPNet	FRCNN	VGG-16	44.9/43.4
CDN	FRCNN	VGG-16	49.3/44.9
ours	FRCNN	VGG-16	**49.8/45.2**
oracle	FRCNN	VGG-16	66.9
SST-AL	FCOS	-	51.8/45.6
CFA	FCOS	VGG-16	49.0/43.2
CFA	FCOS	ResNet-101	51.2/45.0
ours	FCOS	VGG-16	**54.6/48.5**
ours	FCOS	ResNet-101	54.1/46.5
oracle	FCOS	VGG-16	72.3
oracle	FCOS	ResNet-101	71.3

- The "oracle" results indicate that we remove the discriminators in our network and then train and evaluate it on the target domain.

- **Ablation Study**

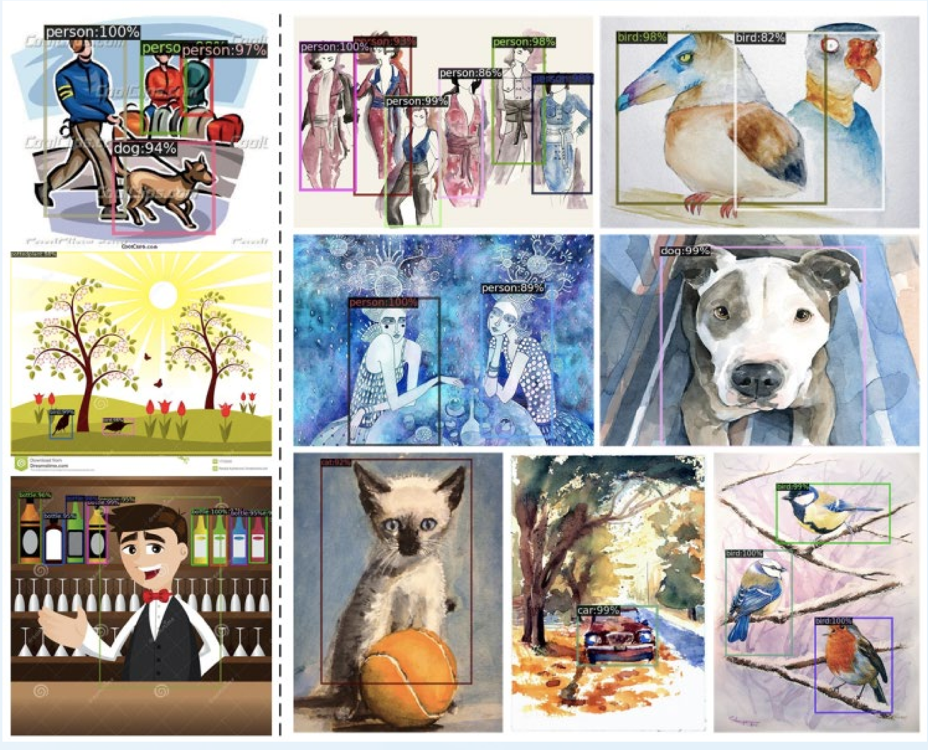
method	mAP	AP^S	AP^M	AP^L
CFA	36.0	8.3	36.7	61.6
ours(w/o all)	36.8	7.2	37.7	64.1
ours(w/o category-level dis.)	39.3	8.7	40.5	64.4
ours(w/o gated fusion)	41.3	8.5	39.1	70.6
ours(w/ all)	43.6	10.1	43.1	72.5
ours(w/ average fusion)	42.1	11.5	40.7	68.9
ours(w/ conv fusion)	41.5	11.2	40.1	71.5
ours(w/ gated fusion)	43.6	10.1	43.1	72.5

discriminator	baseline	D^{cen}	D^{grp}	D^{cls}	D^{cat}(ours)
mAP	39.3	40.5	40.7	41.1	**43.6**

- The proposed omni-scale gated fusion and category-level discriminator reduce false positives and negatives for object detection in adaptive domains.

- **Visualization**

PASCAL VOC ->Clipart **PASCAL VOC ->Watercolor**



Conclusion

- **Contribution**
  - Multi-granularity alignment framework
  - Omni-scale gated fusion
  - Novel category-level discriminator
- **Summary**
  - Applicability of the multi-granularity alignment framework on different detectors
  - Effectiveness of our framework on different domain adaption scenes

Code