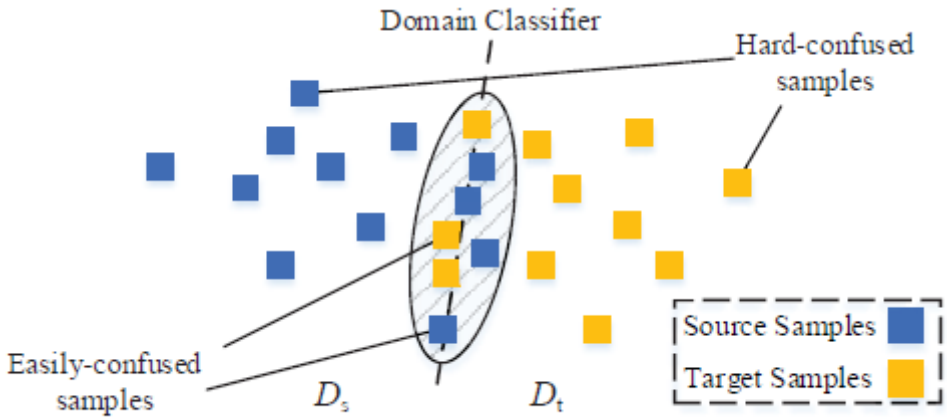
Domain Adaptation Object Detection

遗留问题

- Multi-adversarial Faster-RCNN for Unrestricted Object Detection

  WGRL 的motivations：放松容易混淆的样本，同时对难以混淆的样本进行惩罚，从而更好的domain confusion



$$G_{rev} = -\lambda(d \cdot p + (1-d)(1-p))G \qquad (4)$$

源域的概率是 p, 目标域的概率是$1-p$

相关论文

Adapting Object Detectors via Selective Cross-Domain Alignment

CVPR2019

CODE

Introduction

- 传统的方法是努力将图像**整体对齐**，而对象检测从本质上讲则专注于可能包含感兴趣对象的**局部区域。**
- 因此，本文提出了一种新颖的域自适应方法来处理“where to look”和“how to align”中的问题
- 关键思想是挖掘“discriminative region”，即与对象检测**直接相关的区域**，并专注于在两个域之间对齐它们
- 提出了一个新的框架，包含 **region mining** 和 **region-level alignment**两部分
  - region mining：找到最重要的local区域
  - region-level alignment：利用源域中的区域候选reweight目标区域候选，从而克服由于缺少目标注释而造成的困难，然后以对抗性方式执行region-level alignment。

Method

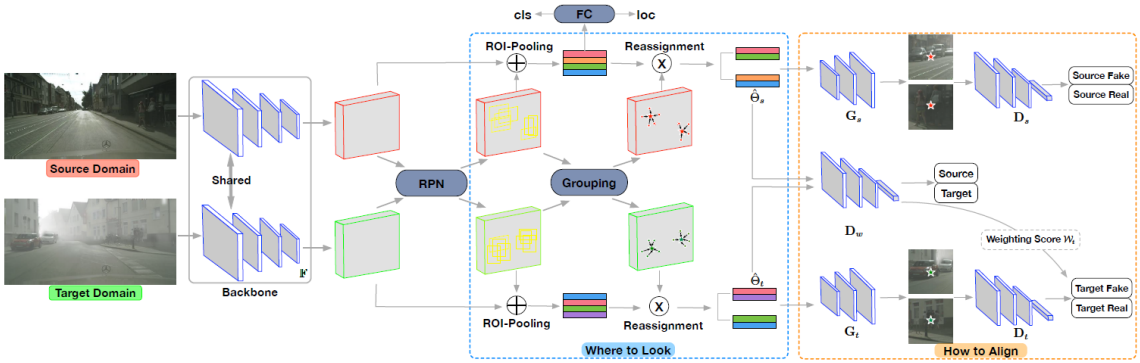


Figure 2. The pipeline of our framework. Two major components, *i.e.* "Where to Look" and "How to Align" are illustrated with two dashed rectangles. For the first component, an ROI-based grouping strategy is designed to mine the discriminative regions for two domains. We display the grouping procedure with cluster number = 2 (Note that ★ and ★ denote the centroids of clusters). For the second one, our model performs the adjusted region-level alignment using generators (G_s and G_t), discriminators (D_s and D_t) and weighting estimator (D_w). We use Faster R-CNN as the detection model ($\mathbb{F}$) which consists of the backbone, RPN and head part. (**Best viewed in color**)

Framework

- 任务是训练一个检测器，该检测器可以利用源域和目标域中的数据很好地泛化到目标域
- 具体来说，希望获得在两个域中均能work的**域不变特征表示**
- 作者提出了一种基于**区域补丁**（region patches）的选择性适应框架
- 基本思想是引入一个附加模块，**根据特征重建图像补丁**，然后在源域和目标域中**对齐重建的补丁**
  - G_s, G_t是patch生成器
- 在训练过程中，该模块可以通过反向传播引导学习特征，从而缩小域之间的差距
- 训练后，不再需要对齐模块，inference的时候只用到检测部分

Region Mining

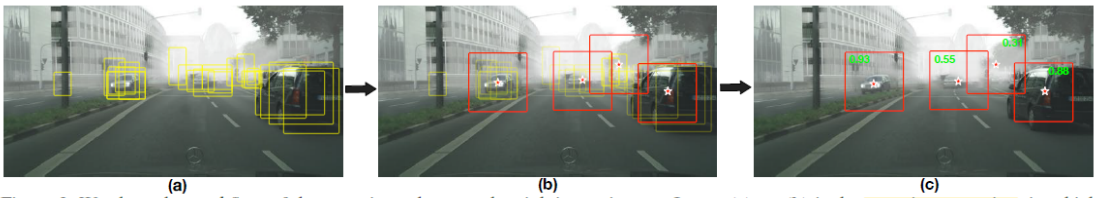
首先通过分组（Grouping）来识别覆盖感兴趣物体对象的重要区域，接着通过重新分配RoI特征来导出这些区域的表征

Grouping

- 我们希望得到大小固定的区域，以便后续操作，但是proposals是具有任意大小的，而且RPN所生成的proposals是包含噪声的
- RPN输出了N_{reg}个region proposals，格式是$\{c_x, c_y, w, h\}$，其中c_x, c_y是中心点坐标。
- 作者利用k-Means聚类方法对这些中心点进行聚类，把这些proposals分成K组，每组的中心作为选择的区域的中心。因为预先设定好区域大小，所以选择的区域已经固定了

Feature Reassignment

- 在第k个region proposals的分组中，有m_k个区域，每个区域的RoI特征的维度是d
- 通过拼接，得到RoI特征的矩阵$\Theta_k \in \mathbb{R}^{m_k \times d}$
- m_k 是变化的，我理解是为了方便计算，需要固定m_k的大小
- 采取的策略是，作者预先设定了一个值m，当m_k的值大于m时，取前m个，小于时，复制现有特征，最终形成矩阵为$\hat{\Theta}_k \in \mathbb{R}^{m \times d}$



(a) (b) (c)

Figure 3. We show the workflow of the grouping scheme and weighting estimator. Image **(a)** → **(b)** is the grouping operation, in which yellow rectangles are the region proposals and red squares denote the discriminative regions. ★ is the centroid of cluster (denoted by Ψ) and also the center of discriminative region. **(b)** → **(c)** is the process of weighting estimation. The green numbers denote the scores of discriminative regions from target domain. (**Best viewed in color**)

Adjusted Region-level Alignment

Region-Level Adversarial Alignment

- 首先，两个生成器G_s和G_t分别在源域和目标域从重要区域的特征$\hat{\Theta}_k$中重构出图像块
- 接着，用两个判别器D_s和D_t分别在源域和目标域中判断真样本和假样本

$$\mathcal{L}_{adv}(\mathbb{F}, G, D) = \mathcal{L}_{adv,D_s} + \mathcal{L}_{adv,D_t} + \mathcal{L}_{adv,G_s} + \mathcal{L}_{adv,G_t} + \mathcal{L}_{adv,\mathbb{F}}. \quad (1)$$

$$\mathcal{L}_{adv}(\hat{\Theta}, \mathcal{P}; G, D) = \mathbb{E}[\log D(\mathcal{P})] + \mathbb{E}[\log(1 - D(G(\hat{\Theta})))], \quad (2)$$

- $L_{adv,D_s}\, L_{adv,D_t}\, L_{adv,G_s}\, L_{adv,G_t}$ 是域内损失,还有交叉域损失$L^s_{adv,F}\, L^t_{adv,F}$, 即把fake source输入到判别器$D_t$中，要求识别成real target;把fake target输入到判别器D_s中，要求识别成real source
- 对齐源域特征和目标域特征，使得模型提取到的特征没有域偏差

Weighting Estimator

- 因为目标域没有ground truth，RPN在目标图像上的region proposals不能覆盖到感兴趣的物体，特别是在训练早期。
- 因此，作者重新对源域的ground truth bounding box重新排序，这可以在目标域中为物体定位提供有用的参考。
- 作者引入一个评估器，根据region proposal与源域的匹配程度来向目标域的区域赋予权重
- 做法是，输入$K * m * d$的特征，判别器D_w 判别源域还是目标域的特征，经过sigmoid输出$K * m$, 然后取均值得到K个值即$\mathcal{W}_t \in \mathbb{R}^K$，表示$K$个regions的权重

最终loss：

Total Objective Function. By incorporating the weighting score $\mathcal{W}_t$, the total optimization of adjusted adversarial alignment can be formulated as:

$$\min_{\mathbb{F},G,D_w} \max_{D} \mathcal{L}_{dec}(\mathbb{F}) + \mathcal{W}_t \cdot \mathcal{L}_{adv}(\mathbb{F}, G, D) + \mathcal{L}_w(D_w), \quad (5)$$

where $\mathcal{L}_{dec}$ is the loss for detection task, *i.e.*, $\mathcal{L}_{dec} = \mathcal{L}_{cls} + \mathcal{L}_{loc}$. $\mathcal{L}_{cls}$ is the cross-entropy loss and $\mathcal{L}_{loc}$ denotes the smooth L1 loss. With the constraint of adjusted region-level adversarial alignment, the training process will encourage domain-invariant features through back-propagation.

Experiment

1. Cityscapes（source）→ Foggy Cityscapes（target）

Methods	person	rider	car	truck	bus	train	motorbike	bicycle	mAP
Source-only	29.7	32.2	44.6	16.2	27.0	9.1	20.7	29.7	26.2
Chen *et al.* [4]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
Ours-Type1(K=2;m=256;512×512)	33.6	37.5	47.8	23.1	39.2	15.2	**29.3**	**34.7**	32.6
Ours-Type2(K=4;m=128;256×256)	**33.9**	**39.7**	**49.7**	21.3	**39.4**	21.9	27.6	34.6	33.5
Ours-Type3(K=8;m=64;128×128)	33.5	38	48.5	**26.5**	39	**23.3**	28	33.6	**33.8**

Table 2. Results of domain adaptation for object detection from Cityscapes to Foggy-Cityscapes (normal → foggy).

2. SIM 10k （source）→ Cityscapes（target）

Methods	*car* AP
Source-only	33.96
Chen *et al.* [4]	38.97
Ours-Type1(K=2;m=256;512×512)	41.97
Ours-Type2(K=4;m=128;256×256)	42.70
Ours-Type3(K=8;m=64;128×128)	**43.02**

Table 3. Results of detection adaptation from synthetic data to real-world data.

3. KITTI（source）→ Cityscapes（target）

Methods	*car* AP
Source-only	37.4
Chen *et al.* [4]	38.5
Ours-Type2(K=4;m=128;256×256)	41.9
Ours-Type3(K=8;m=64;128×128)	**42.5**

Table 4. Results of cross camera adaptation from Kitti dataset to Cityscapes dataset.

思考

- region-level的对齐直觉感受上来说，对于目标检测任务是更合理的，而且从效果上来说确实提升很大
- 但是引入了GAN，不知到训练难度如何

Cross-domain Detection via Graph-induced Prototype Alignment

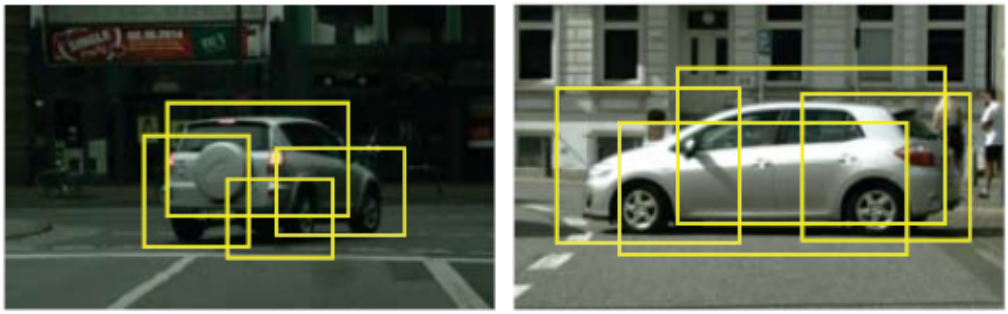
CVPR 2020

CODE

Introduction

- 之前的工作利用独立或分组的区域proposal在local实例级别上对齐源域和目标域，存在两个问题
  - 由于目标域缺少监督信号，导致生成的区域通常与实例相偏移

- 一个实例的representation对于一个类来说是不够的，类内的实例的representation应该是**多模态**的（规模、方向）



(a) (b)

Figure 1. Two vehicles and corresponding region proposals from the Cityscapes [6] dataset which serves as target domain. These two vehicles reflect multi-modal information, *e.g.* distinct scale and orientation, and the generated region proposals contain incomplete information of them.

- 此外，类不平衡问题导致训练过程中不同类之间的域适应过程不一致，导致性能降低
- 为了解决以上问题，提出Graph-induced Prototype Alignment（GPA），图诱导的原型对齐框架，包含两个组件
  - 基于图的区域聚合

    同时考虑定位和候选尺寸关系图用于在实例级别聚合特征，所有实例的重要特征可以被聚合
  - 置信度引导合并

    包含各种实例的多模态信息，通过**prototype representation**来体现，可以利用多模态信息互补，每个类都可以被更好的表达
- 此外，考虑到类间不均衡的的问题，通过 class-reweighted contrastive loss来协调，使得样本缺乏的类被分配更高的权重

Method

$S=\{(x_i^S,y_i^S)\}_{i=1}^{N^s}$ 源域, 用 N_S 表示，服从 P_S 分布；$T=\{(x_j^T)\}_{j=1}^{N^T}$目标域用 N_T 表示，服从 P_T 分布

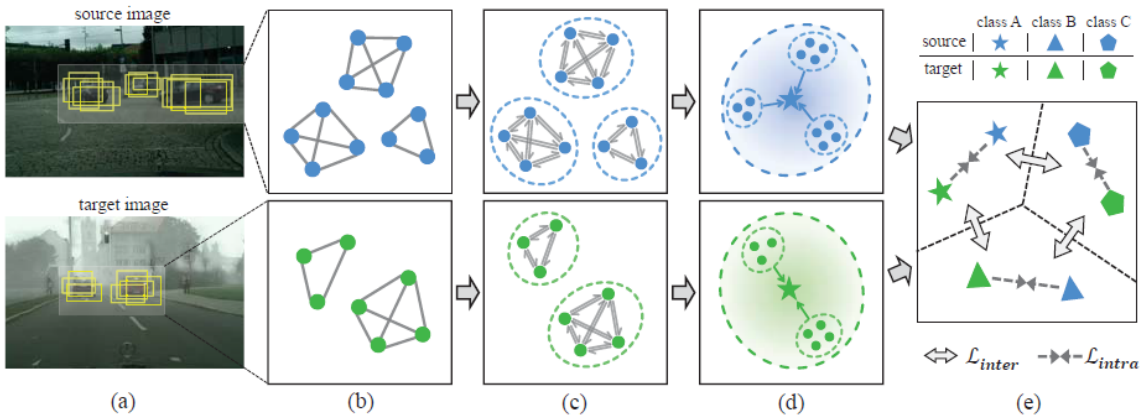


Figure 2. **Framework overview.** (a) Region proposals are generated. (b) Constructing the relation graph on produced region proposals. (c) More accurate instance-level feature representations are obtained through information propagation among proposals belonging to the same instance. (d) Prototype representation of each class is derived via confidence-guided merging. (e) Performing category-level domain alignment through enhancing intra-class compactness and inter-class separability.

- 图诱导的原型矫正（对齐两个域的原型，每个实例的关键信息由图聚合，每个类有每个类的原型）

- 候选区域生成
- 构建关系图

  使用RPN生成的候选区域创建图$G=(V,E)$, V 表示一系列对应于 N_p 的候选区域的图顶点，E 表示图的边，即候选区域的关系，用关系矩阵 A 进行建模。直觉上，空间上近的更可能是一个目标的，因此要分配更大的权重，计算如下：

  $$\mathbf{A}_{i,j}=\exp\left(-\frac{\|o_i-o_j\|_2^2}{2\sigma^2}\right)$$

  其中，o_i,o_j 表示第 i,j 候选区域的中心点, σ 是标准差。

  但是，如下图所属，仅仅用距离表示，是不合理的，应该考虑二者的IoU，使用的计算方法：（两种方法的对比在后续章节）

  $$\mathbf{A}_{i,j}=\mathrm{IoU}(r_i,r_j)=\frac{r_i\bigcap r_j}{r_i\bigcup r_j}$$

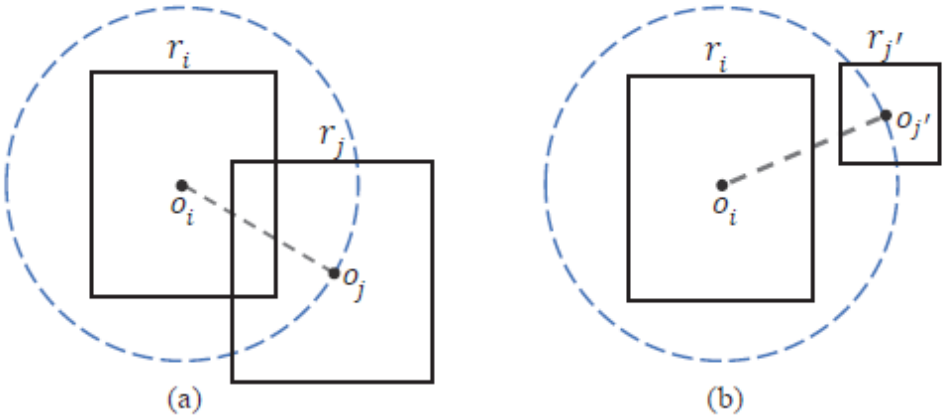


Figure 3. Region proposal r_i interacts with another two region proposals, r_j and $r_{j'}$, with different sizes.

- 基于图的区域整合

  由于检测框的偏移，区域候选往往出现在ground truth周围，导致带有单个候选区域的物体表征不准确。所以为了提取实例级别特征表征，属于某个实例的区域候选框应该被整合。

  $$\widetilde{\mathbf{F}}=\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{F}\ \widetilde{\mathbf{P}}=\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{P}$$

  D 代表对角矩阵，F 是候选区域，$P\in\mathbb{R}^{N_p\times N_c}$ 是分类置信度

  通过相邻proposal之间的信息传播实现实例级信息的表达更加精确。

- 置信度引导合并

  目前的特征已经整合到实例级别，想把不同实例反映出的多模态信息整合到原型中。在合并生成原型的时候，采用候选得分为每个类加权。

  $$c_k=\frac{\sum_{i=1}^{N_p}\widetilde{\mathbf{P}}_{ik}\cdot\widetilde{\mathbf{F}}_i^T}{\sum_{i=1}^{N_p}\widetilde{\mathbf{P}}_{ik}},\qquad(5)$$

  c_k 是第 k 个类的原型

- 类别级域对齐
  - 缩小不同域同一类别的距离（类内），$\mathcal{L}_{intra}$
  - 同时还提出了类间距离 $\mathcal{L}_{inter}$

- 类别不均衡
  - 在 domain adaption中， domain adaption过程对于不同的类别是不平衡的

- 受到Focal loss的影响，对于样本稀缺的类别要赋予高的权重
- 在训练阶段，样本多的类别会更有效的训练和对齐，获得更高的置信度。在一系列候选区域中，选特定类别最高的置信度，用于计算权重

$$p_k = \max_{1 \leqslant i \leqslant N_p} \{\widetilde{\mathbf{P}}_{ik}\},$$

$$\alpha_k = \begin{cases} (1-p_k)^\gamma & \text{if } p_k > \frac{1}{N_c} \\ 0 & \text{otherwise} \end{cases},$$

p_k 表示类别k 中最大的置信度， γ 用于控制不同类权重的参数

- 损失

$$\mathcal{L}_{intra}(\mathcal{S},\mathcal{T}) = \frac{\sum_{i=0}^{N_c} \alpha_i^{\mathcal{S}} \alpha_i^{\mathcal{T}} \Phi(c_i^{\mathcal{S}}, c_i^{\mathcal{T}})}{\sum_{i=0}^{N_c} \alpha_i^{\mathcal{S}} \alpha_i^{\mathcal{T}}}, \tag{8}$$

$$\mathcal{L}_{inter}(\mathcal{D},\mathcal{D}') = \frac{\sum_{0 \leqslant i \neq j \leqslant N_c} \alpha_i^{\mathcal{D}} \alpha_j^{\mathcal{D}'} max(0, m - \Phi(c_i^{\mathcal{D}}, c_j^{\mathcal{D}'}))}{\sum_{0 \leqslant i \neq j \leqslant N_c} \alpha_i^{\mathcal{D}} \alpha_j^{\mathcal{D}'}}, \tag{9}$$

$$\mathcal{L}_{da} = \mathcal{L}_{intra}(\mathcal{S},\mathcal{T}) + \frac{1}{3}\left(\mathcal{L}_{inter}(\mathcal{S},\mathcal{S}) + \mathcal{L}_{inter}(\mathcal{S},\mathcal{T}) + \mathcal{L}_{inter}(\mathcal{T},\mathcal{T})\right), \tag{10}$$

where $\Phi(x, x') = ||x - x'||_2$ calculates the Euclidean distance between two prototypes, and $\{c_i^{\mathcal{S}}\}_{i=0}^{N_c}$, $\{c_i^{\mathcal{T}}\}_{i=0}^{N_c}$ denote the prototypes of source and target domain. $\mathcal{D}$ and $\mathcal{D}'$ represent two domains from which pairs of prototypes belonging to different categories are taken. m is the margin term which is fixed as 1.0 in all experiments. In the total domain adaptation loss $\mathcal{L}_{da}$, all pairwise relations between two domains' prototypes are considered.

Φ 是欧式距离计算

- 两阶段 domain 对齐

$\mathcal{L}_{\text{det}} = \mathcal{L}_{cls}^{RPN} + \mathcal{L}_{\text{loc}}^{RPN} + \mathcal{L}_{cls}^{RCNN} + \mathcal{L}_{\text{loc}}^{RCNN}$

Experiment

Table 1. Experimental results (%) of *Normal to Foggy* cross-domain detection task, Cityscapes → Foggy Cityscapes.

Methods	person	rider	car	truck	bus	train	motorcycle	bicycle	mAP
Source-only	26.9	38.2	35.6	18.3	32.4	9.6	25.8	28.6	26.9
DA [5]	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0
DivMatch [18]	31.8	40.5	51.0	20.9	41.8	34.3	26.6	32.4	34.9
SW-DA [38]	31.8	44.3	48.9	21.0	43.8	28.0	28.9	35.8	35.3
SC-DA [55]	**33.8**	42.1	52.1	**26.8**	42.5	26.5	29.2	34.5	35.9
MTOR [2]	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
GPA (RPN Alignment)	32.5	43.1	53.3	22.7	41.4	40.8	29.4	36.4	37.4
GPA (RCNN Alignment)	33.5	44.8	52.6	26.0	41.2	37.6	29.8	35.2	37.6
GPA (Two-stage Alignment)	32.9	**46.7**	**54.1**	24.7	**45.7**	**41.1**	**32.4**	**38.7**	**39.5**

消融实验

Table 4. Ablation study on different manners to construct relation graph. ("ED": Euclidean distance, "LP": learnable parameter.)

ED	IoU	LP	*car* AP
			45.0
✓			46.1
✓		✓	43.2
	✓		47.6
	✓	✓	43.6

思考

- 也是从 region-level 出发
- 也用到了原型的思想