Cross Domain Object Detection by Target-Perceived Dual Branch Distillation

Mengzhe He[1,3], Yali Wang*[1,6], Jiaxi Wu[5], Yiru Wang[2],
Hanqing Li[2], Bo Li[2], Weihao Gan[2,4], Wei Wu[2,4], Yu Qiao†[1,4]
[1] ShenZhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2] SenseTime Research [3] University of Chinese Academy of Science [4] Shanghai AI Laboratory, Shanghai, China
[5] Beihang University [6] SIAT Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society
{hemz,yl.wang,yu.qiao}@siat.ac.cn, wujiaxi@buaa.edu.cn
{lihanqing,libo,wuwei}@senseauto.com, {wangyiru,ganweihao}@sensetime.com

1. Fourier Domain Image Transfer

$D^s = (x_i^s, y_i^s)$ is a source dataset, where x^s is an image and y^s is an annotation contains bounding boxes and categories information. Similarly $D^t = (x_i^t, y_i^t)$ is a target dataset, where the ground truth annotations are absent. The domain transfer module in [5] are described below.

Let $\mathcal{F}^A, \mathcal{F}^P : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W\times 3}$ be the amplitude and phase components of the Fourier transform F of an RGB image, i.e., for a single channel image x we have:

$$F(x)(m,n)\sum_{h,w} x(h,w)e^{-j2\pi(\frac{h}{H}m+\frac{w}{W}n)}, j^2=-1, \quad (1)$$

which can be implemented efficiently using the FFT algorithm in [1]. Accordingly, $\mathcal{F}^{-1}$ is the inverse Fourier transform that maps spectral signals (phase and amplitude) back to the image space. We assume the center of the image is (0, 0), a mask is defined as following:

$$M_\beta(h,w) = \mathbf{1}_{(h,w)} \in [-\beta H : \beta H, -\beta W : \beta W], \quad (2)$$

then the Fourier domain image transfer can be formalized as

$$x^{s\to t} = \mathcal{F}^{-1}\left(\left[M_\beta \circ \mathcal{F}^A\left(x^t\right) + \left(1-M_\beta\right)\circ \mathcal{F}^A\left(x^s\right), \mathcal{F}^P\left(x^s\right)\right]\right), \quad (3)$$

where the low frequency part of the amplitude of the source image $F^A(x^s)$ is replaced by that of the target image x^t. Then, the modified spectral representation of x^s, with its phase component unaltered, is mapped back to the image $x^{s\to t}$, whose content is the same as x^s, but will resemble the appearance of a sample from D^t. By computing the (Fast) Fourier Transform (FFT) of each input image, and replacing the low-level frequencies of the x^t into the x^s before reconstituting the image for training via the inverse FFT (iFFT), we can get a labeled image with target-like domain style. Smaller β will render the image $x^{s\to t}$ similar with source image x^s. Larger β will make the image $x^{s\to t}$ approach the target image x^t, but also exhibits visible artifacts. We choose $\beta = 0.1$ for our three adaptation scenarios.

* Equal contribution. † Corresponding author.

2. Position Embedding

In this section, we introduce the methods to get geometry similarity between two proposals in detail. Basically, it contains the following two steps. We denote $\mathbf{B}_i$ and $\mathbf{B}_j$ are bounding box prediction of proposal i and j. First, one can compute 4-dimensional relative geometry feature between $\mathbf{B}_i$ and $\mathbf{B}_j$, which denoted as:

$$pos = \left(\log\left(\frac{|x_i - x_j|}{w_i}\right), \log\left(\frac{|y_i - y_j|}{h_i}\right), \log\left(\frac{w_j}{w_i}\right), \log\left(\frac{h_j}{h_i}\right)\right)^T, \quad (4)$$

where x and y are the center coordinates and w and h are the width and height of bounding boxes. Then cosine and sine functions are used to transform this feature as relative position embedding,

$$PE_{(pos,2i)} = sin(pos/1000^{2i/d_{model}}) \quad (5)$$

$$PE_{(pos,2i+1)} = cos(pos/1000^{2i/d_{model}}). \quad (6)$$

Second, an extra FC layer $\mathcal{U}$ is used to project relative position embedding into a scalar weight, which refers to geometry similarity between $\mathbf{B}_i$ and $\mathbf{B}_j$. Additionally, as shown in Eq.7 zero trimming in the max function is used to restrict position comparison between proposals that have geometric relations with high confidence.

$$\mathbf{U}_{i,j} = \max\{0, \mathcal{U}(\mathbf{B}_i, \mathbf{B}_j)\}, \quad (7)$$

Figure 1. Car detection results of different methods for S→C experiments.

Table 1. Position Embedding Effectiveness

Position Embedding	C→ F	S→ C	C→ B
Without	48.3	62.9	43.3
With	49.2	63.3	43.9

Table 2. Head Number Influence

Head Number	C→ F	S→ C	C→ B
L= 4	48.2	63.4	43.8
L=8	47.7	62.5	43.9
L=16	49.2	63.3	43.9

Table 3. Results of CE loss.

Arch	C→ F	S→C
vgg	41.9	50.4
r50	49.1	64.5

3. Ablation and Analysis

Position Embedding. We do ablation studies to see the effectiveness of position embedding. Results in Table 1 show that position embedding is needed in our network. The performance drops slightly when without the position embedding.

Head Number. For our cross attention module MHPCA, we try different head numbers. Table 2 shows that the head number slightly affects the result. Meanwhile, there are also different performance differences on each adaptation scenarios. We set head number=16 by default for our MHPCA module, as it performs well on all datasets.

For the focal loss. Table 3 shows our TDD experiment results with cross entropy. We only show the results of S→C and C→F. The results are comparable with focal loss and still surpass sota.

Visualization We show some detection results of Faster [3], GPA [4], UBT [2] and our TDD on S→C scenario in Figure 1. Only the common category car is reported. Our TDD gives more accurate predictions.

4. Codes and Models

Our code is based on the open-source implementation UBT [2] and the main codes are available in the annex. The detailed environment configuration and our models will be released afterwards.

References

[1] Matteo Frigo and Steven G Johnson. Fftw: An adaptive software architecture for the fft. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 3, pages 1381–1384. IEEE, 1998. 1

[2] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *International Conference on Learning Representations*, 2021. 2, 3

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 2

[4] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12352–12361, 2020. 2

[5] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4084–4094, 2020. 1