Unbiased Faster R-CNN for Single-source Domain Generalized Object Detection Supplementary Material

Yajing Liu[1,2,3], Shijun Zhou[1,2,3], Xiyao Liu[1,2], Chunhui Hao[1,2], Baojie Fan[4], Jiandong Tian[1,2*]
[1]State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences
[2]Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences [4]Nanjing University of Posts and Telecommunications
{liuyajing,zhoushijun,liuxiyao,haochunhui,tianjd}@sia.cn, jobfbj@gmail.com

1. Methodology

1.1. Connection between causality and our method

In our method, the non-causal factors are scenes and non-discriminative object attributes, which lead to attention bias and prototype bias in feature space. Our method is to eliminate the two biases by **learning generalized invariance features robust to the change of non-causal factors** in unseen domains. However, the training data is biased and can't cover a rich set of these non-causal factors. Thus the GLT is proposed to reduce **data bias**. The causal attention learning (**attention debias**) and causal prototype learning (**prototype debias**) modules are designed to learn invariance features from a rich data distribution with diverse non-causal factors. Following the causal theory [1] , the probabilistic invariance constraint $\mathcal{L}_{exp}$ of the predicted results in Eq. (11) is an explicit constraint for causal learning and the feature invariance constraints $\mathcal{L}_{att}$ and $\mathcal{L}_{imp}$in the representation space are implicit ones.

1.2. Reasons of using Dice loss with binary significance maps instead of using MSE loss with attention maps

The binary significance map provides a good representation of the activated and inactivated regions. And the MSE loss constrains the difference in attention values more. For this task, **it is sufficient that the activated regions** are consistency with dice loss though there are differences in the attention values. I think the MSE loss on attention maps is a hard constraint and the dice loss on the significance maps is a soft constraint. Besides, we conduct experiments to analyze the impact of **constraining the attention values with MSE loss** and the mAP on Night-Clear scene decreases by 1.03%.

Methods	Night-Clear	Day-Foggy
Baseline	11.93	8.47
+GLT	7.25	5.18

Table 1. L_1 distance to F_3

Methods	Bus	Bike	Car	Motor	Person	Rider	Truck
UFR w/o $\mathcal{L}_{prot}$	5.16	5.90	5.10	4.86	5.88	6.34	4.95
UFR w/ $\mathcal{L}_{prot}$	3.69	3.17	2.94	4.15	2.72	3.11	4.02

Table 2. L_1 distance between p_i^c and p_{avg}^c

2. Experiments

2.1. More Implementation Details

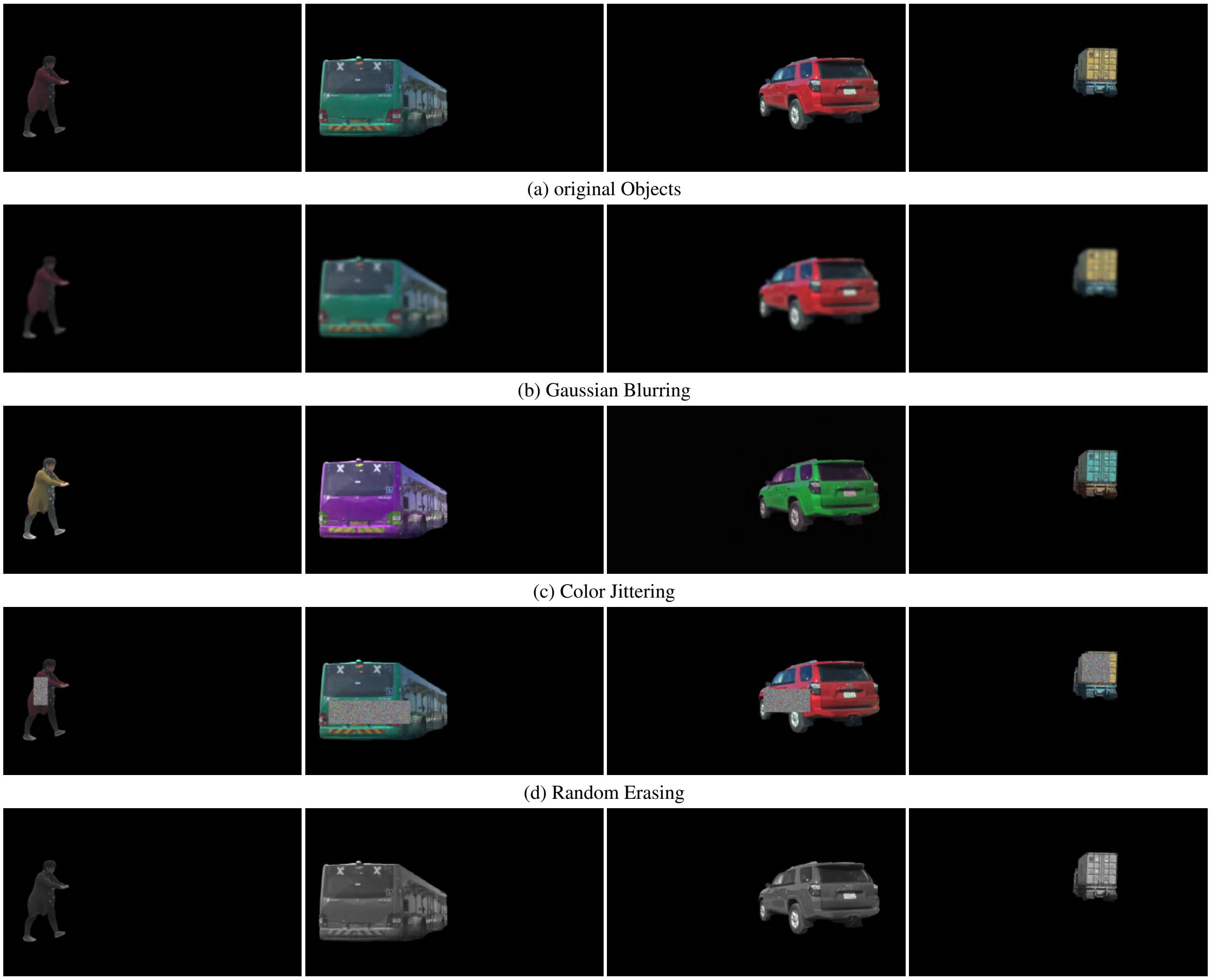
We use Detectron2 [5] on a 24GB GeForce RTX 3090Ti to implement our method. During training, the temperature τ in $\mathcal{L}_{imp}$ is set to 0.2. Besides, the details of the local transformation strategies are as follows:

- **Gaussian Blurring**: We blur the object using a random selected square Gaussian kernel from the size of [23, 27, 29, 31, 33] with standard deviation of 0, as shown in Fig. 1(b).
- **Color Jittering**: We randomly change the brightness, contrast, saturation and hue of an image by a random uniform offset, as demonstrated in Fig. 1(c).
- **Random Erasing**: We randomly select a rectangle region in an object and erase its pixels with random values, as shown in Fig. 1(d).
- **Grayscale**: We randomly apply grayscale on the object, as shown in Fig. 1(e).

2.2. Further analysis of the effectiveness in data debias, attention debias and prototype debias

For data debias, to validate the effectiveness of the GLT module in bridging the domain gap with unseen target domains, we compare the $\boldsymbol{L_1}$ **distance** of features F_1, F_2 generated by **baseline** and **baseline+GLT** respectively with

[*] Corresponding author



(a) original Objects

(b) Gaussian Blurring

(c) Color Jittering

(d) Random Erasing

(e) Grayscale

Figure 1. Visualization examples of the local transformation strategies.

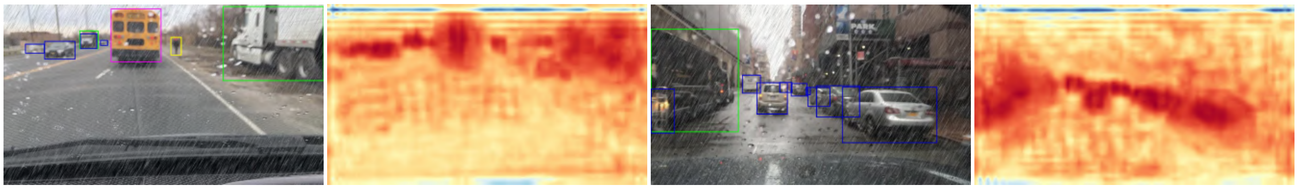


Figure 2. Visualization of the detection results and attention maps on the Dusk-Rainy scene without L_{att}.

feature F_3 generated by model fine tuned in target domain in Table 1. The results indicate that the F_2 of baseline+GLT is closer to F_3, indicating that the GLT module is effective in data debias.

For prototype debias, we sample target images and feed into our **UFR model with** $\mathcal{L}_{prot}$ constraint and **w/o** $\mathcal{L}_{prot}$ constraint respectively to produce the prototypes p_i^c of **category** $\boldsymbol{c}$ in five domains, and then compute the average prototype across five domains for each category c :

$$p_{avg}^c = \frac{1}{5}\sum_{i=1}^{5} p_i^c, \tag{1}$$

where i is the domain number and $i = 1...5$. Then we compute the averaged $\boldsymbol{L_1}$ **distance** d^c for each category between prototype of each domain p_i^c and the average prototype p_{avg}^c to compare the **concentration degree of prototypes** w/ and

τ	0.07	0.2	0.3	1.0
mAP (%)	40.2	40.8	40.5	39.9

Table 3. Results of τ analysis

w/o $\mathcal{L}_{prot}$ following:

$$d^c = \frac{1}{5}\sum_{i=1}^{5}|p_i^c - p_{avg}^c|. \tag{2}$$

The results are shown in Table 2. The results demonstrate that the model can generate more concentrated prototypes of the same category in different domains with our designed $\mathcal{L}_{prot}$, indicating the prototype debias ability of our model.

The effectiveness of attention debias is reflected from the experiment in paper (Fig. 8). We further demonstrate the attention maps generated by our UFR without L_{att} in Fig. 2 for comparison.

2.3. Hyperparameter analysis

We analyze the impact of temperature τ in $\mathcal{L}_{imp}$. The contrastive loss is widely used and we follow the common setting to select τ from [0.07, 0.2, 0.3, 1.0]. We analyze the impact of τ on Night-Clear scene and the results are demonstrated in Table 3. The results show that when the value of τ is small, it has a relatively slight effect on the results, and when τ is increased to 1.0, the results show a significant decrease. We take the value of τ as 0.2 based on this analysis.

3. Discussion on the adoption of SAM[2] in single-domain generalization tasks

In this task, we leverage the powerful segmentation capabilities of SAM [2] to produce object masks of training data. However, the use of SAM [2] may give rise to controversy about whether it violates the single-domain generalization setting. We think that it doesn't violate the single domain setting, for that we don't reach other data or leverage the SAM [2] for testing results. It is just a tool to obtain object masks. The process can be realized with the help of arbitrary segmentation models or even extracted manually. However, we use a more accurate large model to realize it. In addition, with the development of foundation models, there has been a trend of how to leverage them for efficiency gains in various tasks. These models are also used in other cross-domain works, such as the CLIP-Gap [4] in our comparison experiments, which used a CLIP [3] model for domain augmentation.

References

[1] Peter Bühlmann. Invariance, Causality and Robustness. *Statistical Science*, 35:404–426, 2020. 1

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3

[4] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *CVPR*, pages 3219–3229, 2023. 3

[5] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1