Dual Prototype-driven Objectness Decoupling for Cross-Domain Object Detection in Urban Scene Supplementary Materials

1 Implementation Details

We describe more details of various experiment settings in the datasets as we used in the main paper. As demonstrated in the main paper, we utilized the Faster R-CNN [8] with VGG-16 [11] as the default configuration, following the experimental settings [2] as default. Furthermore, we conducted further experiments to demonstrate the flexible scalability of our DuPDA with different backbone, input resolution, and baseline.

Different backbone: In the main paper, we used VGG-16 as the default backbone. However, to demonstrate that our DuPDA can achieve consistent performance regardless of the backbone, we employ ResNet models [3] pretrained on ImageNet [4]. Following the previous studies [9, 13], in the main paper, we use ResNet-50 for weather adaptation using the Cityscapes dataset as denote in Table 3, and ResNet-101 for weather adaptation using the BDD100K dataset as shown in Table 2-(a).

Different input resolution: As depicted in Table 3 of the main paper, we followed the setup in [6] which amplified the input resolution of training and testing scales to enhance weather adaptation using the Cityscapes dataset. The default configuration in our main paper set the shorter dimension of the input image to 600 pixels, as aligned with previous studies [10]; for this specific experiment, we increased it to 800 pixels.

Different baseline: In our main paper, we used the default baseline reported in [10] to performed in overall experiment. Furthermore, to demonstrate the consistency of our DuPDA, we conducted experiments by combining our proposed framework with the mean teacher framework [12], which has recently gained attention in the field of UDA-OD. As shown in Table 1 and Table 2-(b) of the main paper, we utilized the mean teacher framework and achieved competitive results compared to previous works which followed the process described in [1, 5].

2 Error Analysis

To confirm the efficacy of our DuPDA, we compared the highest confidence detection with the baseline [10] in weather adaptation on the Cityscapes dataset. Following the protocols of previous studies [2, 14, 15], we categorized detection results into three groups: **Correct** (IoU with GT ≥ 0.5), **Mislocalization** ($0.3 \leq$ IoU with GT < 0.5), and **Background** (IoU with GT < 0.3). Note that, we selected the top-k predictions within each category, where k equals the number of ground-truth bounding boxes. Subsequently, the average values for each category were computed. Compared to the baseline [10], Fig. 1 clearly shows that DuPDA significantly improves correct detection

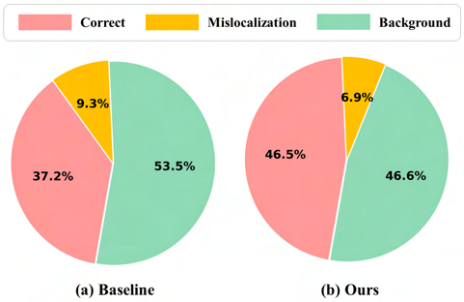


Fig. 1: Error analysis of highest confident detection results compared to the baseline [10].

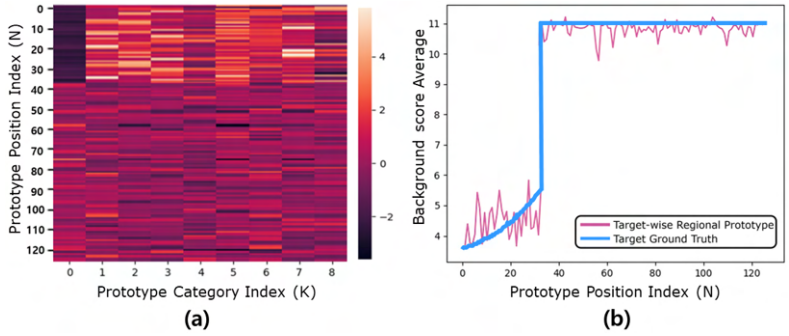


Fig. 2: Visualization of target ROIs extracted from a late iteration result by using DuPDA for training. **(a)** Cumulative average and normalized results of ROIs per category. **(b)** Background classification score distribution of regional prototype, presented alongside the target ground truth.

rates (shown in red portion) and reduces the proportion of background, which indicates as false positive rates (shown in green portion). Furthermore, the mislocalizations (shown in orange portion) are reduced by 2.4% relative to the baseline. The results indicate that DuPDA accurately detects objects and reduces false alarms, contributing to its improved performance in various UDA-OD scenarios.

3 Analysis of Target ROIs trained by DuPDA.

Fig. 2 illustrates the visualization of target ROIs trained via DuPDA. (a) depicts the cumulative mean of the target ROIs, which is trained by DuPDA across various categories of objects and background, subsequently undergoing normalization per category. The 0-th index on the x-axis, indicating the background, is trained to be emphasized toward the higher indices of the y-axis (index $>$ 37). In contrast, the remaining indices on the x-axis that correspond to the foreground are trained to concentrate on the lower indices of the y-axis (index $\leq$ 37). This trend further corroborates the results shown in Fig. 4 of the main paper, indicating that DuPDA successfully guiding the ratio of foreground and background regions using our objectness decoupled loss to unlabeled target ROIs. (b)

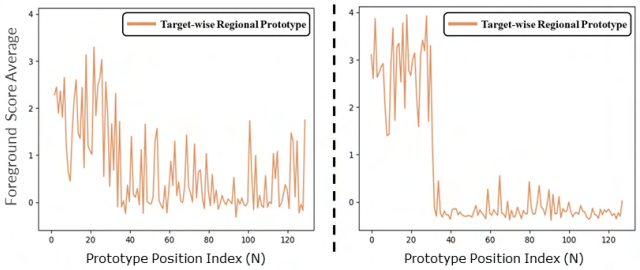
Table 1: Analysis of background refinement loss **(a)** and ablation of our moving alignment **(b)**.

Methods	Only H_T	Random	**Exclusion**
mAP	42.1	41.9	**44.2**

(a) Different strategies for boundary mismatch

Case	without R/CMA	without RMA	without CMA	**R/CMA**
mAP	42.2	43.1	42.9	**44.2**

(b) Different combination using RMA and CMA.



(a) Early Iteration **(b) Late Iteration**

Fig. 3: Foreground classification score distribution in the regional prototype of target domain for the weather adaptation on the Cityscapes. **(a)** Early training iteration. **(b)** Late training iteration.

visualizes the portion of the regional prototype corresponding to the background. This is based on the late iteration result from Fig. 4 in the main paper, along with the target ground truth. When training the target domain using DuPDA without labels, the trained model can closely follow the trend of the actual target labels.

In summary, the DuPDA is training as we intended, which differentiates the foreground and background by leveraging categorical and regional prototypes in unlabled target domain.

4 Analysis on Background Refinement Loss.

When calculating the background refinement loss, we employ both regional prototype P_{R_s} and target ROI scores Ψ_T, partitioned by boundaries H_{P_R} and H_T respectively. However, these boundaries may not always align, producing discrepancies in the range which are classified as background. To address this, we trialed the following 3 methods:

1. Calculating the loss solely based on the extracted H_T from the target ROI scores, disregarding the extracted H_{P_R} from the regional prototype. **(only H_T)**
2. Randomly selecting the specific regional prototype to fill the lacking part, if H_{P_R} < H_T. **(Random)**
3. Applying the background refinement loss exclusively to the overlapping region between thresholds, as described in the main paper Eq. (6). **(Exclusion)**

As shown in Table 1-(a), we adopted the exclusion strategy (3). This strategy calculates the background refinement loss solely for ROIs with distinct boundary definitions, achieved better performance compared to other strategies. This indicates that it is preferable to focus on regions likely to be classified as background while excluding

Table 2: Comparison of the different foreground loss functions: Applying the refinement loss to foreground (Soft Label) vs. our proposed foreground attraction loss (C: Cityscapes, K: KITTI)

Scenario	Refinement loss (Soft Label)	Attraction loss (Ours)
C to F	41.8	**44.2**
C to K	79.2	**81.9**

Table 3: Ablation study for proposed loss functions of DuPDA. **(a)** $\mathcal{L}_{BG}$**:** compare with cross-entropy loss. **(b)** $\mathcal{L}_{FG}$**:** compare with confidence-based pseudo-label generation method.

(a) Different loss of $\mathcal{L}_{BG}$		**(b) Different pseudo-label of $\mathcal{L}_{FG}$**		mAP	
Cross-Entropy	MSE	Confidence	Similarity	C to F	C to K
✓		✓		34.7	71.6
✓			✓	38.5	76.2
	✓	✓		41.3	80.1
	✓		✓	**44.2**	**81.9**

ambiguous regioins. In contrast, strategy (1) posed a risk of misclassifying target ROIs into incorrect foreground categories. Meanwhile, strategy (2) caused confusion for the detector due to the random changes in the regional prototype used for comparing target ROIs at each iteration. Both factors contributed to a decrease in performance.

5 Additional Analysis of Objectness Decoupling.

Fig. 3 depicts the average distribution of all foreground categories in the target-wise regional prototype. The x-axis represents the prototype's indices (N=128), while the y-axis denotes the average of foreground classification scores. During the early iterations, the results show an indistinct separation between the foreground and background regions. However, with the progression of training through our DuPDA, the distinction becomes more pronounced, enhancing the separation between the foreground (index $\leq$ 37) and the background (index $>$ 37). Compared with the visualization of the background region distinction in Fig. 4 of the main paper, it is evident that there is a contrasting trend, with larger values appearing at the front (index $\leq$ 37). These results suggest that our DuPDA helps the model to recognize both foreground and background regions by focusing on training each region separately.

6 Additional Ablation Studies.

This section provides additional ablation studies that were not included in the main paper due to space limitation.

6.1 Effect of Moving Alignments

In Table 1-(b), we present an ablation study on our proposed categorical moving alignment (CMA) and regional moving alignment (RMA), both of which are based on EMA.

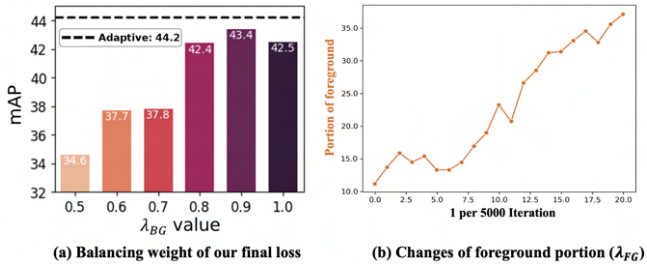


Fig. 4: Analysis of adaptive loss adjustment. **(a)** Comparing our adaptive weight λ_{BG} $(1 - \lambda_{FG})$ with fixed weights. **(b)** Distribution of foreground proportions (y-axis) across iterations (x-axis) directly related to λ_{FG} and λ_{BG}.

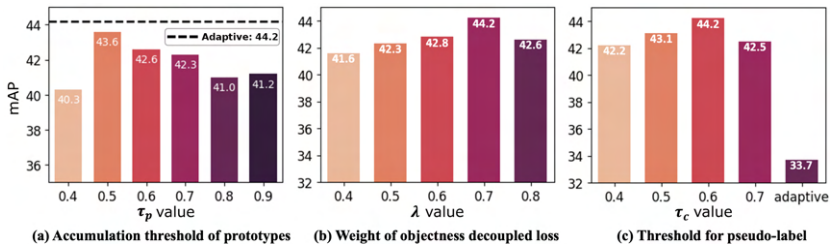


Fig. 5: Analyses of hyperparameter sensitivity. **(a)** Comparing adaptive threshold τ_p with fixed values. **(b)** Different balancing weights λ in our objective function with $(\mathcal{L}_{Det}$ & $\mathcal{L}_{DA}) \leftrightarrow (\mathcal{L}_{FG}$ & $\mathcal{L}_{BG})$. **(c)** Different accumulation threshold τ_c for both categorical and regional prototypes.

These alignments were used to generate the final categorical and regional prototypes from the source- and target-wise categorical and regional prototypes. The results show that the use of both RMA and CMA achieved the best performance, and using RMA alone performed better than using CMA alone. This highlights that the proposed moving alignment method has a significantly greater effect not only on the categorical prototypes, which were often generated by previous UDA-OD methods, but also on the regional prototypes. Note that without CMA or RMA, we replaced them with simply averaging the ROIs to generate the final prototypes.

6.2 Ablation on Objectness Decoupled Loss.

In this section, we show the effectiveness of the proposed loss functions $\mathcal{L}_{FG}$ and $\mathcal{L}_{BG}$. In Table 3-(a), we compared our background refinement loss $\mathcal{L}_{BG}$ with the standard Cross-Entropy loss (CE), which utilizes hard pseudo-labels for background determination in target regions. The result shows that the use of mean square error (MSE) loss outperforms CE when the foreground loss is fixed (see row comparisons $3 \leftrightarrow 5$ and $4 \leftrightarrow 6$ in Table 3). The use of MSE loss, as opposed to use of CE by hard pseudo-labeling, preserves the potential for regions to be classified as foreground, enabling a gradient of likelihood. This reduces misclassification of the background, improving overall detection accuracy.

On the other hand, Table 3-(b) contrasts two pseudo-labeling techniques with $\mathcal{L}_{FG}$: confidence-based pseudo-label generation method using fixed threshold and our similarity-based pseudo-label generation method using adaptive threshold. As shown in the results, our method consistently outperforms the confidence-based alternative when identically applying the background loss, across both UDA-OD scenarios (comparing row 3 $\leftrightarrow$ 4 and comparing between row 5 $\leftrightarrow$ 6 in Table 3). This emphasizes the superiority of our attraction loss using similarity-based pseudo-labels over the refinement loss using soft labels in the foreground region.

Furthermore, Table 2 compares different loss settings for the foreground region using MSE loss with soft labels, which is used for the background refinement loss. The results reveal that our attraction loss, which utilizes similarity-based pseudo-labels, surpasses the performance of the refinement loss using soft labels in both UDA-OD scenarios. This implies that actual categories need to be classified clearly; otherwise, the model may become ambiguous. Consequently, our objectness decoupling loss, using similarity-based pseudo-labels for foreground regions ($\mathcal{L}_{FG}$) and using regional prototypes to calculate the MSE loss for background regions ($\mathcal{L}_{BG}$), proves to be effective in UDA-OD.

6.3 Analysis of adaptive loss adjustment

In the main paper, we adjust $\mathcal{L}_{FG}$ and $\mathcal{L}_{BG}$ based on the number of instances classified as foreground and background within the ROI using λ_{FG} and λ_{BG}. In Fig. 4-(a), we compare fixed versus adaptive weights. Note, for fixed weights, we experimented with setting as $\lambda_{FG} = 1 - \lambda_{BG}$, similar to our approach. The results indicate that adaptive weights (dashed line), which consider the changing boundary of objectness during the training process, surpassed fixed weights. We also show the changes in foreground portion during training that directly influence λ_{FG} and λ_{BG} in Fig. 4-(b). As training progresses, an upward trend is observed, indicating that the detector appropriately adjusts both λ based on its adaptation to domains.

6.4 Additional hyperparameter sensitivity

In the main paper, we initially set λ and τ_c at 0.7 and 0.6, respectively. The parameter λ serves as a balance weight in our objective function of DuPDA, i.e., $\lambda_{FG} = \lambda \cdot (H_T/N)$ and $\lambda_{BG} = \lambda \cdot ((N - H_T)/N)$, comparing it with $\mathcal{L}_{Det}$ and $\mathcal{L}_{DA}$. On the other hand, τ_c is used as a cumulative threshold during the generation of categorical and regional prototypes in both domains. Fig. 5 (b) and (c) present the results derived from varying these parameters. These results reveal that the selected λ and τ_c yield optimal results in our studies, hence their adoption as default settings for all experiments.

Additionally, in Fig. 5-(a), we present different pseudo-label thresholds (τ_p) based on the mean and variance of the similarity map as denoted in the main paper Eq. 1. We found that adaptive thresholds (dashed line) outperformed fixed thresholds. This is because the adaptive threshold considers the magnitude of the overall similarity score, thus mitigating misclassification caused by fixed thresholds.

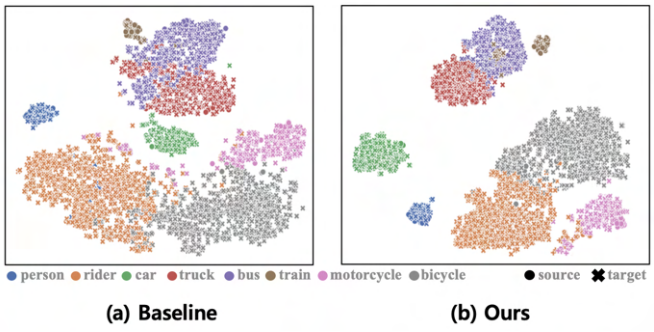


Fig. 6: Visualization of feature comparison by t-SNE. (a) baseline [10], (b) our DuPDA.

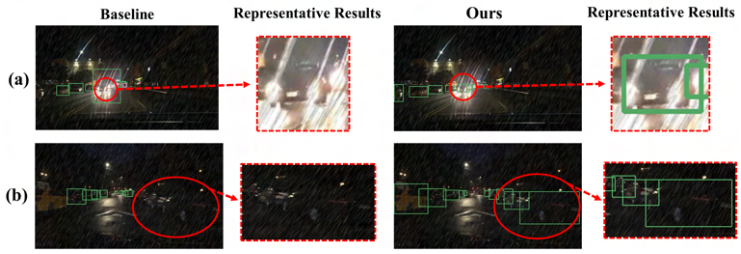


Fig. 7: Analysis of qualitative results in the BDD100K Daytime-sunny to Night-rainy scenario.

7 Detailed Analysis of Qualitative Result

In Fig. 6, we visualize the feature representation using t-SNE [7] in weather adaptation on the Cityscapes dataset. The baseline [10] results show that the clusters of both domains are somewhat grouped, but the object categories overlap without clear separation. However, in our DuPDA results, we observe that not only are the clusters between the source and target domains more compactly gathered, but also the division between the target categories is more distinct. This result indicates that when using DuPDA, the model learns the unique characteristics of objects well, which helps to clearly distinguish the categories of both domains.

We also validate the effectiveness of our DuPDA in improving performance through qualitative results presented in Fig. 7. Compared to the baseline [10], DuPDA demonstrates enhanced detection results for challenging objects affected by light scattering, as shown in Fig. 7-(a). Moreover, Fig. 7-(b) shows clear detection in low-light conditions and for occluded objects. These results highlight the successful transfer of domain-invariant knowledge, including inherent object characteristics, to the target domain. As a result, DuPDA enables robust object detection even in demanding scenarios involving occlusion or variations in lighting conditions in the unlabeled target domain.

8 Additional Qualitative Results

In Fig. 8, 9, we visualize the target domain results for four distinct UDA-OD scenarios. **(a), (b)**: BDD100K Daytime-sunny to Dusk-rainy and BDD100K Daytime-sunny to Night-rainy. **(c), (d)**: Cityscapes to KITTI and Cityscapes to Foggy Cityscapes.

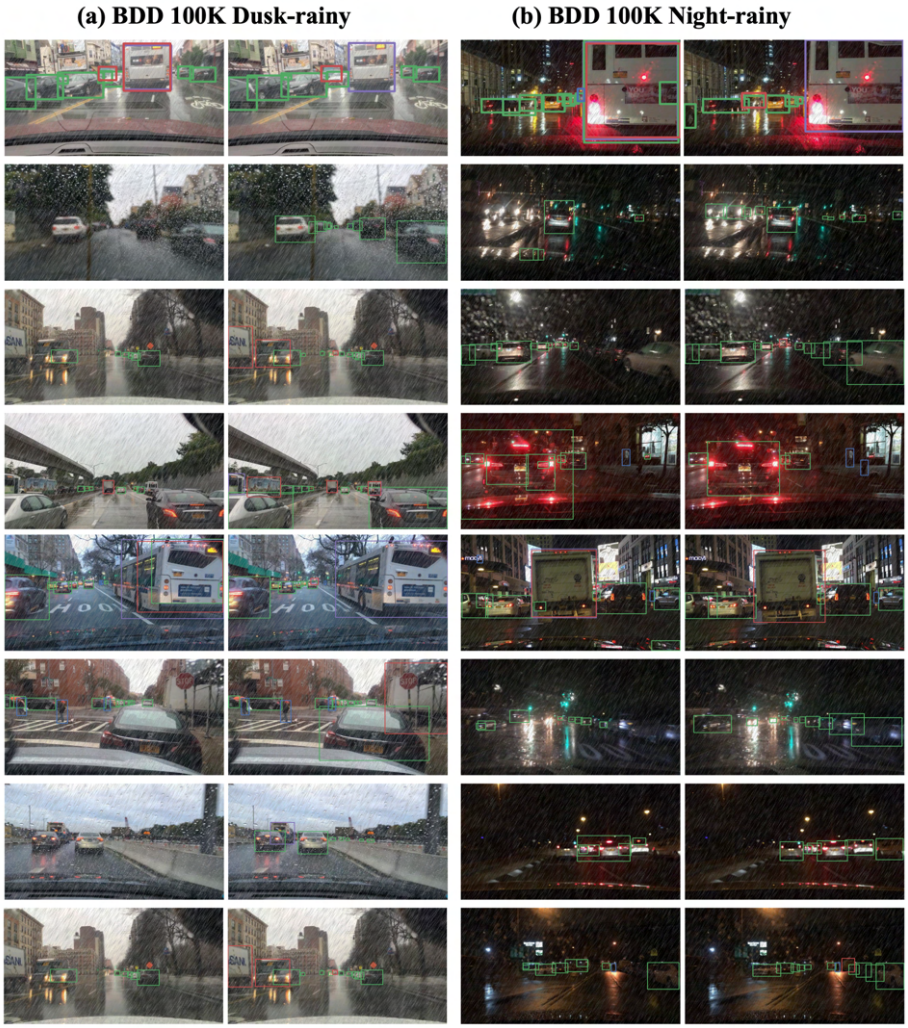


Fig. 8: Further qualitative results from our DuPDA method illustrate the transitions from BDD100K Daytime-sunny to (a) BDD100K Dusk-rainy and (b) BDD100K Night-rainy cases. The comparisons are shown column-wise; from left to right: columns 1 and 3 present the SWDA baseline outcomes, while columns 2 and 4 show the outputs from our DuPDA.

.

Fig. 9: Additional qualitative results employing our proposed DuPDA method are presented for (c) scene adaptation using the KITTI dataset and (d) weather adaptation using the Cityscapes dataset scenarios. The comparisons are displayed column-wise; from left to right: columns 1 and 3 depict the results of the SWDA baseline, while columns 2 and 4 show the results of our DuPDA.

9 Pseudo-code

In this section, we shows PyTorch-like pseudo-code for the objectness decoupling, categorical moving alignment (CMA), and regional moving alignment (RMA).

Pseudo-code PyTorch-like style pseudo-code for Objectness Decoupling.

```
# x_S, y_S: Source input and label   # x_T: Target image
# g_S, g_T: Source and Target classifiers   # f: Baseline model
# cat_PS, reg_PS: Source-wise categorical & regional prototype
# cat_PT, reg_PT: Target-wise categorical & regional prototype
# B_T: Target-wise prototype's boundary of objectness
# B_RP: Regional prototype's boundary of objectness
# L_Det: Source domain detection loss   # S: Similarity
# L_DA: Domain adversarial loss of baseline
# alpha, beta: EMA decay rate   # MSE: Mean Squared Error loss

#compute ROI feature group
h_S, h_T = f(x_S), f(x_T)

# compute ROI score group
z_S, z_T = g_s(h_S), g_t(h_T)
prob_S, prob_T = F.softmax(z_S), F.softmax(z_T)

# generate categorical prototype
cat_PS, cat_PT = CMA(h_S, h_T, prob_S, prob_T, cat_PS, cat_PT)
cat_PS = alpha * cat_PT + (1 - alpha) * cat_PS
# generate regional prototype
reg_PS, reg_PT = RMA(z_S, z_T, prob_S, prob_T, reg_PS, reg_PT)
reg_PS = beta * reg_PT + (1 - beta) * reg_PS

# compute similarity map
similarity = S(cat_PS, h_T)

# compute adaptive threshold for pseudo-label
top1_prob = torch.topk(F.relu(similarity, dim=1), k=1).squeeze()
prob_mean, prob_std = top1_prob.mean(), top1_prob.std()
pth = prob_mean + prob_std

# generate boundary of objectness & pseudo-labels by similarity
y_p = torch.argmax(similarity, dim=0)
_, max_sim = torch.max(similarity, dim=0)
for i in range(len(y_p)):
   if max_sim[i] < pth:
      y_p[i] = 0    #assign background category
      T_B.append(z_T[i])
   else:
      T_F.append(z_T[i])
B_T = len(T_F)
T_F, T_B = torch.stack(T_F, dim=0), torch.stack(T_B, dim=0)
```

```
# compute DuPDA background refinement loss
if B_RP ≤ B_T:
    L_BG = ((N-H)/N) * MSE[T_B[B_T:,:], reg_P[B_T:,:])
else:
    L_BG = ((N-H)/N) * MSE[T_B[(B_RP-H_T):,:], reg_P[B_RP:,:])

# compute DuPDA foreground attraction loss
L_FG = (H/N) * cross_entropy_loss(T_F, y_p)

# compute objectness decoupled Loss
DuPDA_final_loss = L_Det + L_DA + lambda *(L_FG + L_BG)

# optimization step
DuPDA_loss.backward()
update([f.params, g_S.params, g_T.params])
```

Pseudo-code PyTorch-like style pseudo-code for Categorical Moving Alignment.

```
# h_S, h_T: Source and Target ROI feature group
# prob_S, prob_T: Source and Target category probability
# th_S, th_T: Source and Target category threshold
# cat_PS:Source-wise categorical prototype of previous iteration
# cat_PT:Target-wise categorical prototype of previous iteration

#generate categorical prototype for each domain
def CMA(h_S, h_T, prob_S, prob_T, cat_PS, cat_PT):
   max_p_idx_S = torch.argmax(prob_S, dim=1)
   _, max_prob_S = torch.max(prob_S,dim=1)
   max_p_idx_T = torch.argmax(prob_T, dim=1)
   _, max_prob_T = torch.max(prob_T, dim=1)
   for i in range(len(h_S)):    #len(h_S)==len(h_T)==Num of ROIs
     if cat_PS[max_p_idx_S[i]] != None and max_prob_S[i] > th_S:
        cat_PS[max_p_idx_S[i]] += h_S[i]
        cat_PS[max_p_idx_S[i]] /= 2.0
     elif max_prob_S[i] > th_S:
        cat_PS[max_p_idx_S[i]] = h_S[i]
     if cat_PT[max_p_idx_T[i]] != None and max_prob_T[i] > th_T:
        cat_PT[max_p_idx_T[i]] += h_T[i]
        cat_PT[max_p_idx_T[i]] /= 2.0
     elif max_prob_T[i] > th_T:
        cat_PT[max_p_idx_T[i]] = h_T[i]
   return cat_PS, cat_PT
```

Pseudo-code PyTorch-like style pseudo-code for Regional Moving Alignment.

```
# z_S, z_T: Source and Target ROI score group
# prob_S, prob_T: Source and Target category probability
# th_S, th_T: Source and Target category threshold
# reg_PS: Source-wise regional prototype of previous iteration
# reg_PT: Target-wise regional prototype of previous iteration

#generate regional prototype for each domain
def RMA(z_S, z_T, prob_S, prob_T, reg_PS, reg_PT):
   _,max_prob_S = torch.max(prob_S, dim=1)
   _,max_prob_T = torch.max(prob_T, dim=1)
   for i in range(len(z_S)):    #len(z_S)==len(z_T)==Num of ROIs
      if reg_PS[i] != None and max_prob_S[i] > th_S:
         reg_PS[i] = (reg_PS[i] + z_S[i])/2.0
      elif max_prob_S[i] > th_S:
         reg_PS[i] = z_S[i]
      if reg_PT[i] != None and max_prob_T[i] > th_T:
         reg_PT[i] = (reg_PT[i] + z_T[i])/2.0
      elif max_prob_T[i] > th_T:
         reg_PT[i] = z_T[i]
   return reg_PS, reg_PT
```

References

1. Chen, M., Chen, W., Yang, S., Song, J., Wang, X., Zhang, L., Yan, Y., Qi, D., Zhuang, Y., Xie, D., et al.: Learning domain adaptive object detection with probabilistic teacher. arXiv preprint arXiv:2206.06293 (2022)
2. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
5. Li, Y.J., Dai, X., Ma, C.Y., Liu, Y.C., Chen, K., Wu, B., He, Z., Kitani, K., Vajda, P.: Cross-domain adaptive teacher for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7581–7590 (2022)
6. Liu, X., Li, W., Yang, Q., Li, B., Yuan, Y.: Towards robust adaptive object detection under noisy annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14207–14216 (2022)
7. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
8. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
9. Rezaeianaran, F., Shetty, R., Aljundi, R., Reino, D.O., Zhang, S., Schiele, B.: Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In:

Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9204–9213 (2021)
10. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6956–6965 (2019)
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
12. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
13. Xu, M., Wang, H., Ni, B., Tian, Q., Zhang, W.: Cross-domain detection via graph-induced prototype alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12355–12364 (2020)
14. Zhao, L., Wang, L.: Task-specific inconsistency alignment for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14217–14226 (2022)
15. Zheng, Y., Huang, D., Liu, S., Wang, Y.: Cross-domain object detection through coarse-to-fine feature adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13766–13775 (2020)