A. More Implementation Details

In this section, we present more details about the Baseline model. As mentioned in main text (Sec. 3.1), for higher semantic consistency, we adhere to the mainstream practice of aligning features on the source and target domains, at both mid-to-upper layers of the backbone (i.e. image-level) and ROI layer (i.e. instance-level), with the help of Gradient Reversal Layer (GRL) [6]. Concretely, in consistent with [13], for the features output from the last three blocks of VGG16 [14], or last three layers of ResNet101 [8], we feed them into separate discriminators (D1, D2 and D3), their concrete architecture is shown in Tab. A.1 connected via a GRL to determine the domain to which the features belong. After that, three image-level domain adaptation losses are calculated as follows:

\[
L_{\text{img1}}^{\text{da}} = \frac{1}{n_s \cdot H \cdot W} \sum_{i=1}^{n_s} \sum_{w=1}^{W} \sum_{h=1}^{H} D_1(x_i)_{wh}^2 + \frac{1}{n_t \cdot H \cdot W} \sum_{i=1}^{n_t} \sum_{w=1}^{W} \sum_{h=1}^{H} (1 - D_1(x_i)_{wh})^2,
\]

(1)

\[
L_{\text{img2}}^{\text{da}} = \frac{1}{n_s} \sum_{i=1}^{n_s} L_{\text{ce}}(D_2(x_i'), d_i^s) + \frac{1}{n_t} \sum_{i=1}^{n_t} L_{\text{ce}}(D_2(x_i'), d_i^t),
\]

(2)

\[
L_{\text{img3}}^{\text{da}} = \frac{1}{n_s} \sum_{i=1}^{n_s} L_{\text{fl}}(D_3(x_i''), d_i^s) + \frac{1}{n_t} \sum_{i=1}^{n_t} L_{\text{fl}}(D_3(x_i''), d_i^t),
\]

(3)

where \(x_i, x_i'\) and \(x_i''\) denotes the features output from the last three blocks of the backbone for the \(i\)-th training image, \(d_i\) indicates the corresponding domain label, and \(n_s\) and \(n_t\) refer to the total number of images within a mini-batch in source and target domains, respectively. Besides, \(L_{\text{ce}}\) suggests the cross-entropy loss, while the \(L_{\text{fl}}\) indicates the focal loss, with \(\gamma\) set to 5 following [11]. Likewise, the alignment of high-level feature patches (ROIs) is also employed. With the discriminator \(D_d\) illustrated in Tab. A.1, the instance-level loss is formally as

\[
L_{\text{ins}}^{\text{da}} = \frac{1}{n_s} \sum_{i=1}^{n_s} L_{\text{ins}}(r_i, d_i^s) + \frac{1}{n_t} \sum_{i=1}^{n_t} L_{\text{ins}}(r_i, d_i^t),
\]

(4)

where \(r_i\) denotes the \(i\)-th ROI and \(d_i\) indicates the corresponding domain label. As for \(L_{\text{ins}}\), we use cross-entropy loss for the Normal-to-Foggy and Cross-Camera scenarios and focal loss for the Real-to-Artistic scenario, with \(\gamma\) being also set to 5.

In conclusion, the overall training objective of Baseline becomes:

\[
L = L_{\text{det}} + \lambda_1 (L_{\text{img1}}^{\text{da}} + L_{\text{img2}}^{\text{da}} + L_{\text{img3}}^{\text{da}} + L_{\text{ins}}^{\text{da}}),
\]

(5)

where \(\lambda_1\) is set to 1.0. Additionally, we concatenate the image-level features processed by previous three discriminators with the high-level ROI representation after FCs, in a manner similar to [11, 13], to realize greater training stability.

B. Additional Ablation Study

For the localization-specific inconsistency alignment module, the effect of different measures of dispersion is further investigated here. To reveal more clearly their impact on the localization branch, we remove the classification branch. The results on the Real-to-Artistic scenario are displayed in Tab. B.1. It showcases that (1) a measure that is closer to the original scale is preferred; (2) L2-norm delivers a more appropriate and precise estimate to behavioral uncertainty among diverse localizers.
between the source and target domains, while the identical label the former case is ideal, since it shares an gap between the true upper bound and the present upper bound specified by \textit{Target Only} on the former benchmark and easily achieves state-of-the-art performance on the latter benchmark.

It is quite different in the \textit{Cross-Camera} scenario. We find that the label shift of the benchmark (KITTI [7] \leftrightarrow Cityscapes) employed in this scenario is dominated by the imbalance in the foreground-background ratio, namely the inconsistency in the average number of objects between the source and target domain data. In fact, the average numbers of instances of Cityscapes and KITTI are 9.1 and 3.8, respectively. This directly leads to two serious problems. On the one hand, we observe that the \textit{Source Only} model undergoes severe overfitting issue during training, which means that we underestimate the lower bound of the benchmark; on the other hand, it imposes higher demands on the cross-domain performance of RPN, and this straightforwardly undermines the effectiveness of the existing mainstream approaches that focus on feature alignment for it.

In summary, two arguments are made. First, existing methods are highly inefficient in coping with label shift. In light of [15], although the execution of domain alignment alone reduces the divergence between domains (the second term in Theorem 1), it leads to arbitrary increases in $\lambda^*$ (the third term in Theorem 1), hence eventually, the target errors of detectors cannot be well-guaranteed. For this reason, taking into account the detectors’ empirical predictions on the target domain, or namely, the behavior of label predictors, is gradually emerging as a necessity. Moreover, compared to classification tasks, the label shift in object detection task is considerably complicated. It is no longer limited to the differences in category proportions, but is more widely distributed in spatial differences in scale, position, etc. of bounding boxes. These two facts drive the proposal of TIA on a different aspect.

Second, in view of the fact that the label shift cannot be well estimated nor truly eliminated, we argue that there is a gap between the true upper bound and the present upper bound specified by \textit{Target Only}, according to [16]. Under such circumstances, the close performance of the domain space between the source and target domains, while the latter one has its label shift diluted due to the scale of the source domain. In this context, it is observed that, our framework exceeds the upper bound indicated by \textit{Target Only} on the former benchmark and easily achieves state-of-the-art performance on the latter benchmark.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute deviation</td>
<td>42.7</td>
</tr>
<tr>
<td>Variance</td>
<td>41.8</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>43.2</td>
</tr>
</tbody>
</table>

Table B.1. Ablation study on different measures of dispersion.

### C. Visualization

We provide some detection results of vanilla detector (\textit{i.e. Source Only} [10]), state-of-the-art adaptive detectors (\textit{e.g.} HTCN [1] and UMT [4]), and our framework TIA. Fig. F.1 illustrates the comparison of detections on the PASCAL VOC [5] \rightarrow Clipart [9] benchmark. It is observed that our proposed TIA outperforms both \textit{Source Only} and \textit{UMT} [4], and produces more accurate detection results, \textit{i.e.}, more foreground objects are identified (Row 1&2), and higher quality bounding boxes are provided along with accurate categorization (Row 3-5). Qualitative results on the Cityscapes [3] \rightarrow Foggy Cityscapes [12] benchmark represented by Fig. F2 also demonstrates the superiority of our TIA. For example, in the first row, for the two cars on the left, the bounding box given by HTCN is relatively off-target, while ours method present more compact boundaries, compared to \textit{Source Only}’s.

### D. Limitations

The discrepancy between source and target domains in the label space, \textit{i.e.}, label shift, substantially affects the design philosophy and severely limits the performance of existing domain adaptive detectors. In this subsection, we will provide in-depth analysis of how label shift limits our TIA for each dataset benchmark.

The benchmarks used in Normal-to-Foggy (Cityscapes [3] \rightarrow Foggy Cityscapes [12]) and Real-to-Artistic (PASCAL VOC [5] \rightarrow Clipart [9]) are essentially appropriate and they allow a good evaluation of the performance of various domain adaptive detectors. Specifically, the former case is ideal, since it shares an identical label space between the source and target domains, while the latter one has its label shift diluted due to the scale of the source domain. In this context, it is observed that, our framework exceeds the upper bound indicated by \textit{Target Only} on the former benchmark and easily achieves state-of-the-art performance on the latter benchmark.

In light of [15], although the execution of domain alignment alone reduces the divergence between domains (the second term in Theorem 1), it leads to arbitrary increases in $\lambda^*$ (the third term in Theorem 1), hence eventually, the target errors of detectors cannot be well-guaranteed. For this reason, taking into account the detectors’ empirical predictions on the target domain, or namely, the behavior of label predictors, is gradually emerging as a necessity. Moreover, compared to classification tasks, the label shift in object detection task is considerably complicated. It is no longer limited to the differences in category proportions, but is more widely distributed in spatial differences in scale, position, etc. of bounding boxes. These two facts drive the proposal of TIA on a different aspect.

### Table A.1. Architecture of discriminators.

<table>
<thead>
<tr>
<th>Discriminator $D_1$</th>
<th>Conv 1 × 1 × 256, stride 1, pad 0</th>
<th>ReLU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Conv 1 × 1 × 128, stride 1, pad 0</td>
<td>ReLU</td>
</tr>
<tr>
<td></td>
<td>Conv 1 × 1 × 1, stride 1, pad 0</td>
<td>Sigmoid</td>
</tr>
<tr>
<td>Discriminator $D_2$ and $D_3$</td>
<td>Conv 3 × 3 × 512, stride 2, pad 1</td>
<td>Batch Normalization, ReLU, Dropout</td>
</tr>
<tr>
<td></td>
<td>Conv 3 × 3 × 128, stride 2, pad 1</td>
<td>Batch Normalization, ReLU, Dropout</td>
</tr>
<tr>
<td></td>
<td>Conv 3 × 3 × 128, stride 2, pad 1</td>
<td>Batch Normalization, ReLU, Dropout</td>
</tr>
<tr>
<td></td>
<td>Average Pooling</td>
<td>Fully connected 128 × 2</td>
</tr>
<tr>
<td>Discriminator $D_4$</td>
<td>Conv 3 × 3 × 512, stride 2, pad 1</td>
<td>ReLU</td>
</tr>
<tr>
<td></td>
<td>Conv 3 × 3 × 128, stride 2, pad 1</td>
<td>ReLU</td>
</tr>
<tr>
<td></td>
<td>Conv 3 × 3 × 128, stride 2, pad 1</td>
<td>ReLU</td>
</tr>
<tr>
<td></td>
<td>Average Pooling</td>
<td>Fully connected 128 × 2</td>
</tr>
</tbody>
</table>
adaptive detectors in the *Cross-Camera* benchmark can be reasonably explained.

**E. Societal Impact**

Domain adaptive object detection is a prevalent visual scene understanding task and we follow the convention experimental setting as in [1, 2, 4, 11]. Hence if the method is used properly, there is no negative social impact.

**F. Code and Dataset License**

Our code is built on open-sourced object detection code with MIT license. As for the datasets, the Cityscapes [3] and its modification Foggy Cityscapes [12] are made freely available to academic and non-academic entities for non-commercial purposes such as academic research, teaching, scientific publications, or personal experimentation; and the PASCAL VOC [5] includes images obtained from the "flickr" website; Clipart [9] is meant for education and research purposes only; in addition, KITTI [7] is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License.

**References**


Figure F.1. Illustration of the detection results on the PASCAL VOC → Clipart benchmark. Compared to Source Only, UMT’s localization performance is worse, while ours is better.
Figure F.2. Illustration of the detection results on the Cityscapes \(\rightarrow\) Foggy Cityscapes benchmark. Our TIA identifies more objects and delivers more accurate bounding boxes.