# Decoupled Adaptation for Cross-Domain Object Detection

Junguang Jiang, Baixu Chen, Jianmin Wang, Mingsheng Long

Tsinghua University

JiangJunguang1123@outlook.com

# Problem

- Object detection in the real world suffers from performance drop due to variance in viewpoints, object appearance, backgrounds, illumination, image quality, *etc*.
- *Domain adaptation* aims to transfer a detector from a source domain $\mathcal{D}_s$, where sufficient training data is available, to a target domain $\mathcal{D}_t$ where only unlabeled data is available.

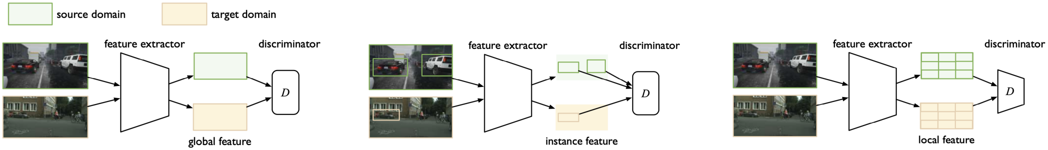

(a) Labeled Source Domain

(b) Unlabeled Target Domain

# Data Challenge

What to adapt in the object detection task is unknown.

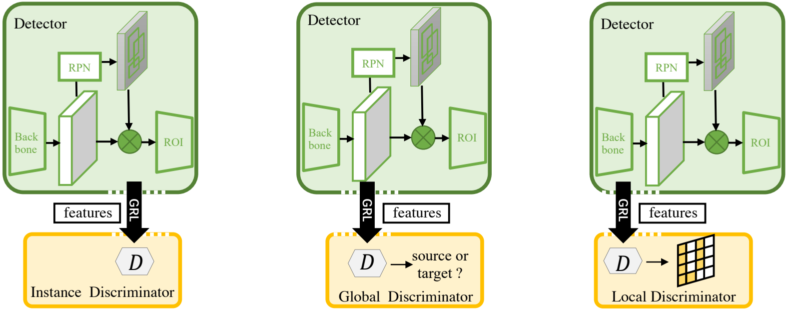
- Global feature adaptation in the image level is likely to mix up features of different objects since each input image of detection has multiple objects.
- Instance feature adaptation in the object level might confuse the features of the foreground and the background.
- Local feature adaptation in the pixel level will struggle when the domains are different at the semantic level.




(c) Global adapt (d) Instance adapt (e) Local adapt

# Architecture Challenge

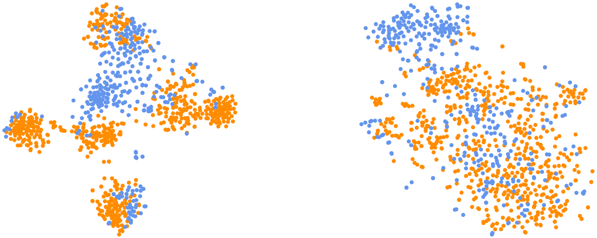
- Introducing domain discriminators and gradient reverse layers into the detector architecture to encourage domain-invariant features might deteriorate the discriminability of detectors.
- The scalability of previous methods to different detection architectures is not so satisfactory.




(f) Instance adapt (g) Global adapt (h) Local adapt

# Task Challenge

- Object detection is a multi-task learning problem, consisting of both classification and localization.
- Yet previous adaptation algorithms mainly explored the category adaptation, and it's still difficult to obtain an adaptation model suitable for different tasks at the same time.



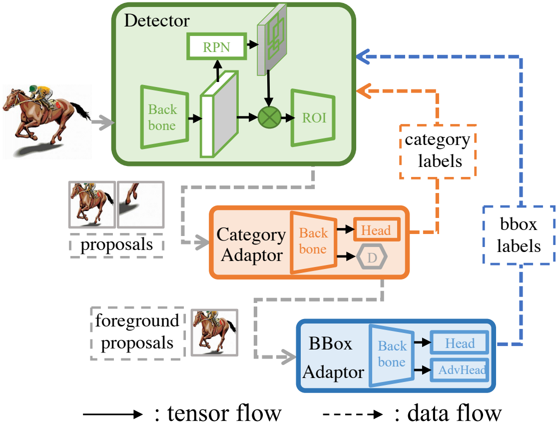
(i) Classification Features (j) Detection Features

# D-adapt Framework

- **Procedures:**
  1. Decouple the original cross-domain detection problem into several sub-problems.
  2. Design domain adaptors to solve each sub-problem.
  3. Coordinate the relationships between different adaptors and the detector.
- **Meanings:** Different parts have
  - independent model parameters,
  - independent input data distributions,
  - independent training losses,

  and are coordinated into some relationships through data rather than gradients
- **Summary:**
  1. Architecture challenge $\Rightarrow$ Decouple the adversarial adaptation from the training of the detector by introducing a parameter-independent category adaptor.
  2. Task challenge $\Rightarrow$ Introduce another bounding box adaptor that's decoupled from both the detector and the category adaptor.
  3. Data challenge $\Rightarrow$ Adjust the object-level data distribution for specific adaptation tasks.

# D-adapt Framework



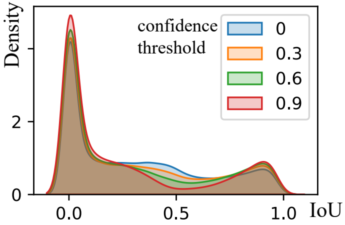

**Algorithm 1** Training Pipeline.

**input** : Source domain $\mathcal{D}_s$ and target domain $\mathcal{D}_t$, number of iterations $T$
**output:** Cross-domain object detector $G^{\text{det}}$

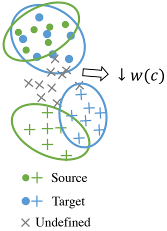
*initialize* the object detector $G^{\text{det}}$ by optimizing with $\mathcal{L}_s^{\text{det}}$
**for** $t \leftarrow 1$ **to** $T$ **do**
    generate proposals $\mathcal{D}_s^{\text{prop}}$ and $\mathcal{D}_t^{\text{prop}}$ for each sample in $\mathcal{D}_s$ and $\mathcal{D}_t$ by $G^{\text{det}}$
    **for** *each mini-batch in* $\mathcal{D}_s^{prop}$ *and* $\mathcal{D}_t^{prop}$ **do**
        train the category adaptor $G^{\text{cls}}$;
    **end**
    generate category label for each proposal in $\mathcal{D}_t^{\text{prop}}$
    generate foreground proposals $\mathcal{D}_s^{\text{fg}}$ and $\mathcal{D}_t^{\text{fg}}$ from $\mathcal{D}_s^{\text{prop}}$ and $\mathcal{D}_t^{\text{prop}}$
    **for** *each mini-batch in* $\mathcal{D}_s^{fg}$ *and* $\mathcal{D}_t^{fg}$ **do**
        train the bounding box adaptor $G^{\text{reg}}$;
    **end**
    generate bounding box label for each proposal in $\mathcal{D}_t^{\text{fg}}$
    train the object detector $G^{\text{det}}$ by optimizing with $\mathcal{L}_t^{\text{det}}$
**end**

# Category Adaptation

- **Objective**: Use labeled source-domain proposals $(\mathbf{x}_s, \mathbf{y}_s^{\text{gt}}) \in \mathcal{D}_s^{\text{prop}}$ to obtain a relatively accurate classification $\mathbf{y}_t^{\text{cls}}$ of the unlabeled target-domain proposals $\mathbf{x}_t \in \mathcal{D}_t^{\text{prop}}$.
- **Data challenge**: The input data distribution doesn't satisfy the low-density separation assumption well $\Rightarrow$ impede the adversarial alignment.
- **Solution**: Use the confidence of each proposal to discretize the input space,
  - When a proposal has a high confidence $\mathbf{c}^{\text{det}}$ being the foreground or background, it should have a higher weight $w(\mathbf{c}^{\text{det}})$ in the adaptation, and vice versa




(k) IoU distribution of proposals




(l) Discretization

# Category Adaptation

- The objective of the discriminator $D$ is,

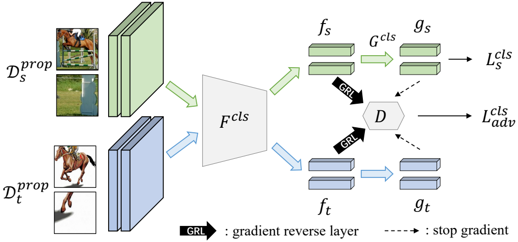
$$\max_{D} \mathcal{L}_{\text{adv}}^{\text{cls}} = \mathbb{E}_{\mathbf{x}_s \sim \mathcal{D}_s^{\text{prop}}} w(\mathbf{c}_s) \log[D(\mathbf{f}_s, \mathbf{g}_s)] + \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}_t^{\text{prop}}} w(\mathbf{c}_t) \log[1 - D(\mathbf{f}_t, \mathbf{g}_t)], \quad (1)$$

where $\mathbf{f} = F^{\text{cls}}(\mathbf{x})$ is the feature and $\mathbf{g} = G^{\text{cls}}(\mathbf{f})$ is the category prediction.

- The objective of the feature extractor $F^{\text{cls}}$ is

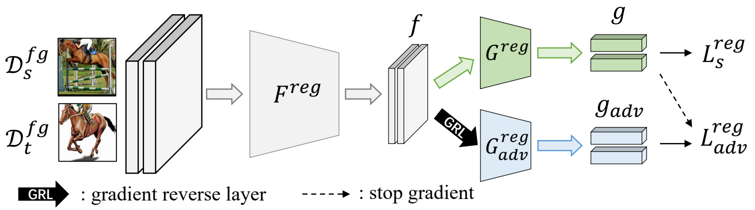
$$\min_{F^{\text{cls}}, G^{\text{cls}}} \mathbb{E}_{(\mathbf{x}_s, \mathbf{y}_s^{\text{gt}}) \sim \mathcal{D}_s^{\text{prop}}} \mathcal{L}_{\text{CE}}(G^{\text{cls}}(\mathbf{f}_s), \mathbf{y}_s^{\text{gt}}) + \lambda \mathcal{L}_{\text{adv}}^{\text{cls}}, \quad (2)$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, $\lambda$ is the trade-off.

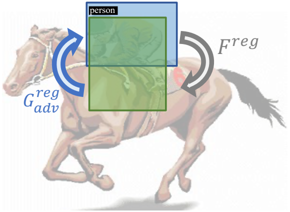


# Bounding Box Adaptation

- **Objective**: Utilize labeled source-domain foreground proposals $(\mathbf{x}_s, \mathbf{b}_s^{\text{gt}}) \in \mathcal{D}_s^{\text{fg}}$ to obtain bounding box labels $\mathbf{b}_t^{\text{reg}}$ of the unlabeled target-domain proposals $\mathbf{x}_t \in \mathcal{D}_t^{\text{fg}}$.
- **Architecture**: A feature generator network $F^{\text{reg}}$ which takes proposal inputs, and two regressor networks $G^{\text{reg}}$ and $G_{\text{adv}}^{\text{reg}}$ which take features from $F^{\text{reg}}$.
- **Method**: Optimize the adversarial regressor network $G_{\text{adv}}^{\text{reg}}$ to maximize its disparity with the main regressor on the target domain while minimizing the disparity on the source domain to measure the discrepancy across domains.




(m) Architecture of the bounding box adaptor




(n) Minimax on IoU

# Quantitative Results

- Experiments show that *D-adapt* achieves state-of-the-art results on four cross-domain object detection tasks and yields 17% and 21% relative improvement on benchmark datasets *Clipart1k* and *Comic2k* in particular.

**Table:** Results from PASCAL VOC to Clipart (ResNet101, 20 categories).

| | aero | bcycle | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | hrs | bike | prsn | plnt | sheep | sofa | train | tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source Only | 35.6 | 52.5 | 24.3 | 23.0 | 20.0 | 43.9 | 32.8 | 10.7 | 30.6 | 11.7 | 13.8 | 6.0 | 36.8 | 45.9 | 48.7 | 41.9 | 16.5 | 7.3 | 22.9 | 32.0 | 27.8 |
| DA-Faster | 15.0 | 34.6 | 12.4 | 11.9 | 19.8 | 21.1 | 23.2 | 3.1 | 22.1 | 26.3 | 10.6 | 10.0 | 19.6 | 39.4 | 34.6 | 29.3 | 1.0 | 17.1 | 19.7 | 24.8 | 19.8 |
| BDC-Faster | 20.2 | 46.4 | 20.4 | 19.3 | 18.7 | 41.3 | 26.5 | 6.4 | 33.2 | 11.7 | 26.0 | 1.7 | 36.6 | 41.5 | 37.7 | 44.5 | 10.6 | 20.4 | 33.3 | 15.5 | 25.6 |
| WST-BSR | 28.0 | 64.5 | 23.9 | 19.0 | 21.9 | 64.3 | 43.5 | 16.4 | 42.0 | 25.9 | 30.5 | 7.9 | 25.5 | 67.6 | 54.5 | 36.4 | 10.3 | 31.2 | 57.4 | 43.5 | 35.7 |
| SWDA | 26.2 | 48.5 | 32.6 | 33.7 | 38.5 | 54.3 | 37.1 | 18.6 | 34.8 | 58.3 | 17.0 | 12.5 | 33.8 | 65.5 | 61.6 | 52.0 | 9.3 | 24.9 | 54.1 | 49.1 | 38.1 |
| MAF | 38.1 | 61.1 | 25.8 | **43.9** | 40.3 | 41.6 | 40.3 | 9.2 | 37.1 | 48.4 | 24.2 | 13.4 | 36.4 | 52.7 | 57.0 | **52.5** | 18.2 | 24.3 | 32.9 | 39.3 | 36.8 |
| SCL | 44.7 | 50.0 | 33.6 | 27.4 | 42.2 | 55.6 | 38.3 | 19.2 | 37.9 | **69.0** | 30.1 | **26.3** | 34.4 | 67.3 | 61.0 | 47.9 | 21.4 | 26.3 | 50.1 | 47.3 | 41.5 |
| CRDA | 28.7 | 55.3 | 31.8 | 26.0 | 40.1 | 63.6 | 36.6 | 9.4 | 38.7 | 49.3 | 17.6 | 14.1 | 33.3 | 74.3 | 61.3 | 46.3 | 22.3 | 24.3 | 49.1 | 44.3 | 38.3 |
| HTCN | 33.6 | 58.9 | 34.0 | 23.4 | **45.6** | 57.0 | 39.8 | 12.0 | 39.7 | 51.3 | 21.1 | 20.1 | 39.1 | 72.8 | 63.0 | 43.1 | 19.3 | 30.1 | 50.2 | 51.8 | 40.3 |
| ATF | 41.9 | **67.0** | 27.4 | 36.4 | 41.0 | 48.5 | 42.0 | 13.1 | 39.2 | 75.1 | 33.4 | 7.9 | 41.2 | 56.2 | 61.4 | 50.6 | **42.0** | 25.0 | 53.1 | 39.1 | 42.1 |
| Unbiased | 30.9 | 51.8 | 27.2 | 28.0 | 31.4 | 59.0 | 34.2 | 10.0 | 35.1 | 19.6 | 15.8 | 9.3 | 41.6 | 54.4 | 52.6 | 40.3 | 22.7 | 28.8 | 37.8 | 41.4 | 33.6 |
| D-adapt | **56.4** | 63.2 | **42.3** | 40.9 | 45.3 | **77.0** | **48.7** | **25.4** | **44.3** | 58.4 | **31.4** | 24.5 | **47.1** | **75.3** | **69.3** | 43.5 | 27.9 | **34.1** | **60.7** | **64.0** | **49.0** |

# Qualitative results

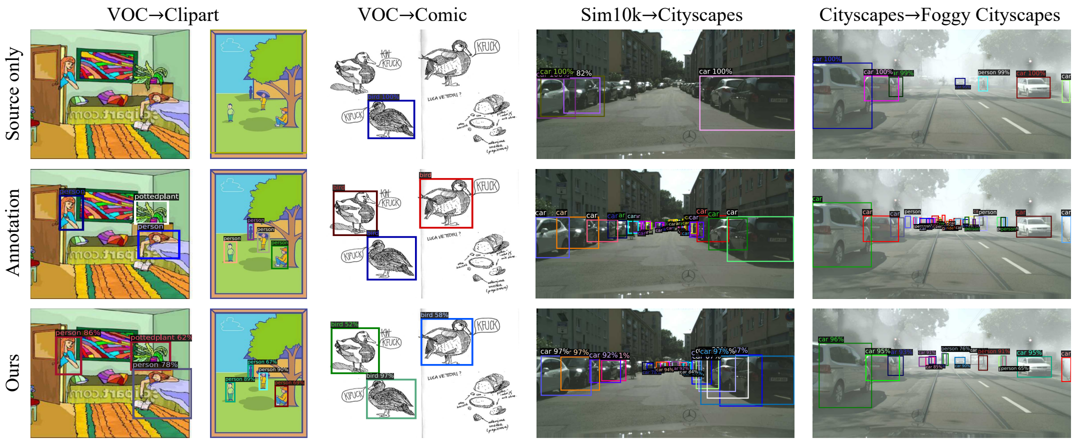



**Figure:** Qualitative results on the target domain.

# Ablation

- Category adaptation. The weight mechanism has the greatest impact, indicating the necessity of the low-density assumption in the adversarial adaptation.
- Bounding box adaptation. Minimizing the disparity discrepancy improves the performance of the box adaptor and bounding box adaptation improves the performance of the detector in the target domain.

**Table:** Ablations on PASCAL VOC to Clipart.

(a) Category adaptation

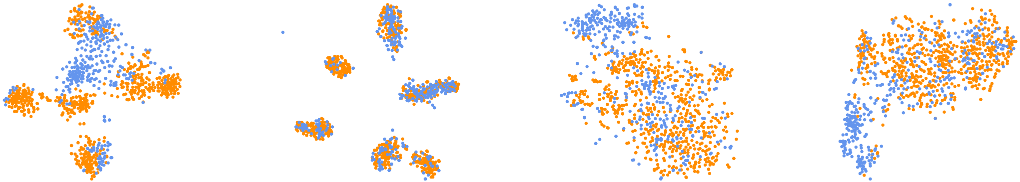
| metric | ours | w/o condition | w/o bg proposals | | | | w/o weight | w/o adaptor |
|---|---|---|---|---|---|---|---|---|
| | | | source<br>target | ✗<br>✗ | ✗<br>✔ | ✔<br>✗ | | |
| $\text{mIoU}^{\text{cls}}$<br>mAP | 38.2<br>43.5 | 36.9<br>41.7 | -<br>- | 36.6<br>41.7 | 33.6<br>38.8 | 25.1<br>36.5 | 17.2<br>33.3 | 12.6<br>28.0 |

(b) Spatial Adaptation

| metric | Ours | w/o DD | w/o adaptor |
|---|---|---|---|
| $\text{mIoU}^{\text{reg}}$<br>mAP | 0.631<br>45.0 | 0.598<br>44.4 | 0.531<br>43.5 |

# Analysis

- *The features of the detector do not have an obvious cluster structure, even on the source domain.*
- The reason is that the features of the detector contain both category information and location information.



(a) Adaptor ($\lambda = 0$) (b) Adaptor ($\lambda = 1$) (c) Baseline (mAP:19.7) (d) Ours (mAP:40.5)

**Figure:** T-SNE visualization of features on task VOC $\rightarrow$ Comic2k (6 classes). (a) and (b) are features from the category adaptor. (c) and (d) are features from the Faster RCNN.

# Discussions

- Our method achieved considerable improvement on several benchmark datasets for domain adaptation.
- D-adapt framework does not introduce any computational overhead in the inference phase, since the adaptors are independent of the detector and can be removed during detection. In actual deployment, the detection performance can be further boosted by employing stronger adaptors without introducing any computational overhead.
- D-adapt does not depend on a specific detector, thus the detector can be replaced by SSD, RetinaNet, or other detectors.
- D-adapt framework can be extended to other detection tasks, e.g., instance segmentation and keypoint detection, by cascading more specially designed adaptors.

# Open-Source Library

- Our code is available at `https://github.com/thuml/Transfer-Learning-Library/tree/dev-tllib/examples/domain_adaptation/object_detection`.
- **TLlib** is an open-source and well-documented library for Transfer Learning and has got over 1.4$k$ stars.
- Supported applications include: classification, segmentation, object detection, re-identification, keypoint detection and so on.