This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3082687, IEEE Transactions on Multimedia

Uncertainty-Aware Unsupervised Domain Adaptation in Object Detection

Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, Yanpeng Cao

***Abstract*—Unsupervised domain adaptive object detection aims to adapt detectors from a labelled source domain to an unlabelled target domain. Most existing works take a two-stage strategy that first generates region proposals and then detects objects of interest, where adversarial learning is widely adopted to mitigate the inter-domain discrepancy in both stages. However, adversarial learning may impair the alignment of well-aligned samples as it merely aligns the global distributions across domains. To address this issue, we design an uncertainty-aware domain adaptation network (UaDAN) that introduces conditional adversarial learning to align well-aligned and poorly-aligned samples separately in different manners. Specifically, we design an uncertainty metric that assesses the alignment of each sample and adjusts the strength of adversarial learning for well-aligned and poorly-aligned samples adaptively. In addition, we exploit the uncertainty metric to achieve curriculum learning that first performs easier image-level alignment and then more difficult instance-level alignment progressively. Extensive experiments over four challenging domain adaptive object detection datasets show that UaDAN achieves superior performance as compared with state-of-the-art methods.**

***Index Terms*—Unsupervised domain adaptation, object detection, adversarial learning, curriculum learning.**

I. Introduction

OBJECT detection aims to locate and recognize objects of interest in images, which has been a longstanding challenge in computer vision research [1]–[6]. With the development of deep convolutional neural networks in recent years, object detection has achieved great progress [7]–[24] over multiple large-scale datasets [25]–[29]. However, existing object detection methods usually experience drastic performance drops when applied to new datasets due to various domain biases in camera settings, environmental illumination, object appearance, etc. Annotating a fair number of samples for each new data collection can alleviate this problem effectively, but it is often prohibitively time-consuming and unscalable while facing various new data.

Unsupervised domain adaptation has been explored extensively [30]–[37] to address domain biases by learning a well-performed model on an unlabeled target domain with supervision of a labeled source domain. Motivated by the domain adaptation theory [38] that the upper bound of target-domain errors could be reduced by minimizing the divergence between source and target domains, most existing domain adaptive object detection methods adopt adversarial learning to minimize the cross-domain discrepancy [39]–[48]. Though these works have achieved impressive performance, they usually suffer from a common constraint of the adversarial learning - it merely considers the alignment of global distributions and may pull well-aligned samples to incorrect classes as illustrated in Fig. 1 (a).

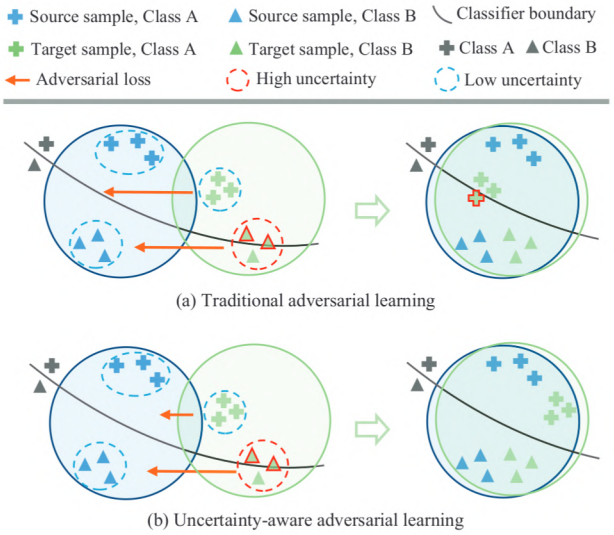

Fig. 1. The proposed uncertainty-aware adversarial learning (UaAL) performs adversarial learning adaptively: Traditional adversarial learning assigns adversarial loss of the same weight to all samples (global alignment), which may misalign well-aligned samples (far from classifier boundary with low uncertainty) to incorrect classes as in (a). UaAL weights adversarial loss adaptively based on the uncertainty of each sample. It can thus protect well-aligned samples from further alignment while focusing on aligning under-aligned samples (near to classifier boundary with high uncertainty) as in (b). Here falsely classified samples are highlighted in red color and the length of orange-color arrows denotes adversarial-loss weights - UaAL assigns smaller weights (shorter arrow in (b)) to well-aligned samples. Best viewed in color.

In this paper, we propose an uncertainty-aware domain adaptation network (UaDAN) that aligns well-aligned and poorly-aligned samples adaptively by considering their uncertainty. With supervised models trained using source samples, well-aligned target samples are often predicted with low uncertainty while poorly-aligned target samples predicted with high uncertainty. We exploit this uncertainty information and achieve uncertainty-aware cross-domain alignment with two innovative designs. The first design is uncertainty-aware adversarial learning (UaAL) as illustrated in Fig. 1 (b). Instead of assigning the same adversarial-loss weight to all target samples equally, UaAL introduces uncertainty to measure an alignment

Corresponding author: Shijian Lu (e-mail: Shijian.Lu@ntu.edu.sg)

D. Guan, J. Huang, A. Xiao and S. Lu are with Singtel Cognitive and Artificial Intelligence Lab for Enterprises, Nanyang Technological University, Singapore.

Y. Cao is with the School of Mechanical Engineering, Zhejiang University, Hangzhou, China.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3082687, IEEE Transactions on Multimedia

score for each sample and adjusts the adversarial-loss weights for well-aligned and poorly-aligned target samples adaptively. In implementation, we exploit entropy [49] to estimate the sample uncertainty and weight adversarial loss of different target samples adaptively (*i.e.*, small weight for well-aligned samples and large weight for poorly-aligned samples).

The second design is uncertainty-guided curriculum learning (UgCL) that aims to optimize the domain adaptation across the two-stage detection pipeline progressively. Specifically, UgCL first aligns target samples at an easier image level for region proposal generation. It exploits entropy to estimate the uncertainty of image-level predictions for uncertainty-aware alignment at the image level. After target samples are well-aligned at the image level (with low uncertainty), UgCL aligns at a harder instance level for final object detection. Here the uncertainty is exploited to guide the instance-level alignment in the similar manner. The two-stage progressive alignment strategy alleviates error accumulation effectively by reducing image-level misalignment and its effects over the later instance-level alignment.

The contributions of this work can be summarized in three aspects. *First*, we propose an uncertainty-aware domain adaptation network that assesses the uncertainty of sample predictions and employs it for adaptive sample alignment effectively. *Second*, we develop an uncertainty-aware adversarial learning method that alleviates misalignment of well-aligned samples by assigning high weights to high-uncertainty samples and low weights to low-uncertainty samples. *Third*, we design an uncertainty-guided curriculum learning technique that aligns samples first at the easier image level and then at the harder instance level progressively, ultimately leading to robust cross-domain alignment.

The remainder of this paper is organized as follows. We first review related works in Section II. Our proposed uncertainty-aware domain adaptation method is then presented in Section III. Extensive experimental results are presented in Section IV, and Section V finally concludes this paper.

II. Related Works

A. Object detection

Existing object detection methods can be broadly divided into two categories: proposal-based and proposal-free. Proposal-free methods consider detection as a bounding box regression problem. For example, YOLO [10] directly predicts detection results by regressing the coordinates of predefined anchors and simultaneously classifying categories. SSD [11] integrates predictions computed from hierarchical networks with multiple respective fields to deal with instances of different scales. Several extensions [12]–[15] have been proposed for more effective proposal-free object detectors. Proposal-based methods first generate region proposals and then classify them for final detection results. For example, R-CNN [7] uses a hierarchical grouping algorithm to extract dense region proposals and then classify these proposals to obtain detection results. Fast R-CNN [8] speeds up the R-CNN by introducing a region-of-interest (ROI) pooling layer to share features across each proposal. Faster R-CNN [9] proposes a more efficient and accurate region proposal network (RPN) to replace the hierarchical grouping applied in Fast R-CNN [8]. Several extensions [16]–[20] present more powerful proposal-based detectors for better detection performance.

Most existing object detection methods require a large amount of labelled training data which often takes time to annotate. This paper presents a domain adaptive detector built upon Faster R-CNN [9] that aims to optimally exploit the training data that are collected and annotated in prior scenes.

B. Domain adaptive detection

Our work is closely related to the area of knowledge transfer [50]–[52] and domain adaptive object detection [39]–[46], which aims to learn a well-performed detector on unlabeled target domain without accessing any annotations in target domain. Quite a number of domain adaptive detectors have been reported. For example, DA [39] presents a domain adaptive detector based on Faster R-CNN to minimize domain bias via adversarial learning at both image and instance levels. MTOR [53] minimizes domain discrepancy by integrating object relations into teacher-student consistency regularization. SWDA [40] presents a powerful cross-domain detection method via strongly aligning local similar features and weakly aligning global dissimilar features. IFAN [42] aligns feature distributions at both image and instance levels in a coarse-to-fine manner. ATF [43] presents a tri-stream Faster R-CNN to alleviate the collapse risk caused by parameter sharing between source and target domains. CTR [44] trains the region proposal network (RPN) and the object detection classifier via collaborative self-training. GPA [54] aligns graph-induced prototype representations in two stages to minimize domain discrepancy. CRDA [46] presents an effective categorical regularization module that improves DA [39] and SWDA [40] consistently.

The aforementioned cross-domain detection methods mainly employ adversarial learning which merely considers alignment of global distributions and may pull well-aligned samples incorrectly. This work instead aims for protecting well-aligned samples by identifying well-aligned and poorly-aligned samples and aligning them separately in different manners.

C. Curriculum learning

Curriculum learning has been widely explored in the past decade [55]–[57]. Bengio *et al.* [55] argues that the generalization of supervised networks could be gradually increased via training from easy examples to harder ones. Kumar *et al.* [56] determines the order of training samples based on the context of non-convex optimization. Recently, curriculum learning has been widely applied in supervised learning [58]–[61] and semi-/weakly-supervised learning [62]–[66]. In the context of cross-domain adaptation, Zhang *et al.* [67] minimizes the domain bias in semantic segmentation via a curriculum-style learning method, where easy tasks are solved first for inferring necessary target-domain properties. Zheng *et al.* [45] coarsely aligns foreground regions in feature space and finely aligns the class prototype distance.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3082687, IEEE Transactions on Multimedia

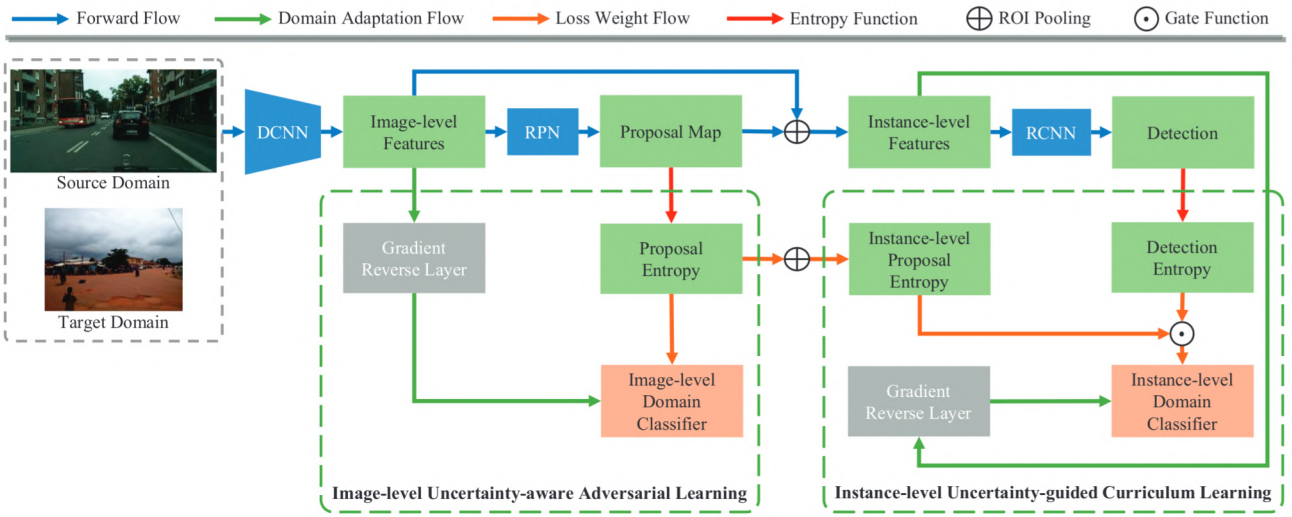

Fig. 2. The framework of our uncertainty-aware domain adaption network (UaDAN): UaDAN adopts Faster R-CNN architecture with a deep convolutional neural network (DCNN), an RPN and a region convolutional neural network (RCNN). RPN is fed with image-level features from DCNN and generates region proposals. RCNN takes instance-level features from a ROI pooling layer as input and produces the final detection. In image-level adaptation, image-level domain classifier is fed with image-level features from a gradient reversal layer (GRL) where uncertainty-aware adversarial loss is weighted by classification entropy of region proposals. In instance-level adaptation, instance-level domain classifier is fed with instance-level features from the GRL and optimized by our uncertainty-guided curriculum learning (UgCL). With UgCL, detection entropy is filtered by instance-level proposal entropy using a gate function and then fed as the loss weight of instance-level domain classifier. Instance-level proposal entropy is computed from proposal entropy map with an ROI pooling layer.

Our work builds on top of the curriculum learning idea by creating a curriculum over subtasks that a domain adaptive algorithm needs to solve. Different from existing works, it introduces uncertainty awareness into curriculum learning to alleviate the side effect of adversarial learning in the harder instance-level alignment subtask.

III. Method

This work focuses on domain adaptive object detection that consists of two sub-tasks in proposal generation and instance detection. We design an uncertainty-aware domain adaptation technique that introduces uncertainty-aware adversarial learning and uncertainty-guided curriculum learning for aligning image-level and instance-level features adaptively and progressively as illustrated in Fig. 2, more details to be described in the following subsections.

A. Problem Definition

We consider a set of source images $X_s \in \mathbb{R}^{H\times W\times 3}$ with object labels $\hat{Y}_s = \{\hat{C}_s, \hat{B}_s\}$, where $\hat{C}_s \in (1, C)^N$ and $\hat{B}_s \in (1, C)^{N\times 4}$ represent object categories C and bounding box coordinates respectively, and a set of target images $X_t \in \mathbb{R}^{H\times W\times 3}$ without labels. Here, H, W, N denote image height, image width, and the class number of objects, respectively. The goal of domain adaptive object detection is to learn a well-performed detector G with access of $\{X_s, \hat{Y}_s, X_t\}$.

Motivated by domain adaptation theory in [38], recent domain adaptive detection methods [39]–[46] adopt Faster R-CNN as backbone and employ traditional adversarial learning (TDA) to minimize the cross-domain discrepancy at image and instance levels. They usually exploit G to distill knowledge from $\{X_s, \hat{Y}_s\}$ by minimizing a supervised loss $\mathcal{L}_{det}$, and learn domain-invariant representations via a minimax game between G and domain classifiers (C_{img} and C_{ins}) under adversarial learning losses ($\mathcal{L}_{img}$ and $\mathcal{L}_{ins}$). The objective function of TDA in object detection is thus a combination of the three losses:

$$\mathcal{L}_{TDA} = \mathcal{L}_{det}(F) + \mathcal{L}_{img}(C_{img}) + \mathcal{L}_{ins}(C_{ins}), \quad (1)$$

Under the guidance of adversarial losses $\mathcal{L}_{img}$ and $\mathcal{L}_{ins}$ in Eq. 1, TDA minimizes cross-domain discrepancy by aligning features at image and instance levels directly. However, such brute-force inter-domain alignment may introduce misalignment. Specifically, TDA may misalign well-aligned samples to incorrect categories as it assigns adversarial loss of the same weight to all samples. We define this issue as a global alignment problem, and design conditional adversarial losses to identify and align well-aligned and poorly-aligned samples separately in different manners.

B. Network Architecture

As illustrated in Fig. 2, the proposed cross-domain object detection network G adopts Faster R-CNN as the backbone. The network G consists of a DCNN, an RPN and an RCNN that focus on deep feature extraction, region proposal generation and bounding box detection, respectively. RPN is fed with image-level features from DCNN to predict region category (*i.e.*, 0 for background and 1 for foreground) and generate candidate bounding boxes via dense anchors. RCNN is fed with instance-level features from a ROI pooling layer to classify the candidate bounding boxes generated by RPN into pre-defined C classes and refine their coordinates.

Source images X_s are fed to the detection network for optimizing G via a supervised loss with Y_s, and simulta-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3082687, IEEE Transactions on Multimedia

neously produce source features (*i.e.*, $F_{img_s} \in \mathbb{R}^{U\times V\times D}$ at image level and $F_{ins_s} \in \mathbb{R}^{M\times M\times K}$ at instance level). Here, $\{U, V, D\}$ and $\{M, M, K\}$ denote the three dimensions (*i.e.*, height, width and depth) of the image-level and instance-level features, respectively. Target images X_t are fed into the network G to generate target features (*i.e.*, $F_{img_t} \in \mathbb{R}^{U\times V\times D}$ at image level and $F_{ins_t} \in \mathbb{R}^{M\times M\times K}$ at instance level) where domain classifiers learn how target features map to source features via image-level and instance-level domain adaptation modules. We adopt the GRL [68] for extracting domain-invariant features via adversarial learning. In implementation, GRL lying between the detection network G and each domain classifier (*i.e.,*, C_{img} at the image level or C_{ins} at the instance level) works by reversing gradients from each domain classifier to the G (in back-propagation). G thus receives the reversed gradients and updates its parameters in an opposite direction, and this guides G to generate domain-invariant features for deceiving the domain classifiers. We followed [39] to construct the image-level and instance-level domain classifiers. Specifically, the image-level domain classifier has two convolutional layers whose dimensions are $1\times1\times1024\times512$ and $1\times1\times512\times1$, respectively. The instance-level domain classifier has three fully-connected layers, where the first two layers have 1024 channels each and the third layer contains 1 channel. To avoid overfitting in the instance-level domain classifier, the dropout rate is set to 0.5 for the first two fully-connected layers.

In image-level domain adaptation, image-level domain classifier C_{img} is fed with image-level features (F_{img_s} and F_{img_t}) from GRL and optimized via an image-level uncertainty-aware adversarial loss $\mathcal{L}_{img}^{ua}$ (weighted by proposal entropy). The proposal entropy map ($E_{P_s} \in \mathbb{R}^{U\times V}$ and $E_{P_t} \in \mathbb{R}^{U\times V}$) is computed from proposal maps ($P_s \in \mathbb{R}^{U\times V\times R}$ and $P_t \in \mathbb{R}^{U\times V\times R}$) in both source and target domains. Here, R denote the anchors of different scales and ratios at each location (u, v). With the uncertainty-aware adversarial loss, our network will focus on aligning under-aligned features with high entropy while keeping well-aligned features with low entropy less affected.

In the instance-level domain adaptation, instance-level domain classifier C_{ins} is fed with instance-level features (F_{ins_s} and F_{ins_t}) from the GRL and optimized by instance-level uncertainty-guided curriculum loss $\mathcal{L}_{ins}^{ug}$ (weighted by the filtered detection entropy). The filtered detection entropy is computed by a gate function $\mathcal{G}$ whose input is detection entropy ($E_{D_s} \in \mathbb{R}^K$ and $E_{D_t} \in \mathbb{R}^K$) and truncation parameter is instance-level proposal entropy ($E_{ins_s} \in \mathbb{R}^{C,K}$ and $E_{ins_t} \in \mathbb{R}^{C,K}$). We apply an ROI pooling layer to generate instance-level proposal entropy (E_{ins_s} and E_{ins_t}) from proposal entropy (E_{P_s} and E_{P_t}). With the uncertainty-guided curriculum loss, instance-level alignment can only be activated when the instance-level proposal entropy is lower than a given truncation threshold. It can ensure more robust alignment by first aligning features at easier image level and then aligning features at harder instance level progressively. The symbols used in this paper are listed in Table I.

TABLE I
GLOSSARY OF SYMBOLS.

Symbols	Description
X_s	The set of source images
$\hat{Y}_s$	The set of source labels
X_t	The set of target images
G	The object detection network
C_{img}	The image-level domain classifier
C_{ins}	The instance-level domain classifier
P_s	The region proposals in source domain
P_t	The region proposals in target domain
D_s	The object detection in source domain
D_t	The object detection in target domain
F_{img_s}	The image-level features in source domain
F_{img_t}	The image-level features in target domain
F_{ins_s}	The instance-level features in source domain
F_{ins_t}	The instance-level features in target domain
E_{P_s}	The proposal entropy in source domain
E_{P_t}	The proposal entropy in target domain
E_{ins_s}	The instance-level proposal entropy in source domain
E_{ins_t}	The instance-level proposal entropy in target domain
E_{D_s}	The detection entropy in source domain
E_{D_t}	The detection entropy in target domain
L_s	The source domain labels
L_t	The target domain labels
O_{img_s}	The image-level source domain predictions
O_{img_t}	The image-level target domain predictions
O_{ins_s}	The instance-level source domain predictions
O_{ins_t}	The instance-level target domain predictions
ξ	The hyper-parameter to activate instance-level alignment

C. Training Objective

The network G is optimized with three loss terms, *i.e.*, a supervised detection loss $\mathcal{L}_{det}$ for distilling knowledge in source domain, an image-level uncertainty-aware adversarial loss $\mathcal{L}_{img}^{ua}$ and an instance-level uncertainty-guided adversarial loss $\mathcal{L}_{ins}^{ug}$ for unsupervised domain adaptation. In the forward processing, we fed the network G with a pair of source and target images (x_s and x_t) and obtain region proposals (p_s and p_t) from RPN, detection results (d_s and d_t) from RCNN, image-level features (f_{img_s} and f_{img_t}) from DCNN, and instance-level features (f_{ins_s} and f_{ins_t}) from an ROI pooling layer which extract local information from image-level features in line with the proposal. Our training objective is to maximize mixed likelihood of source and target features.

1) ***Source-domain supervised learning****:* For source-domain supervised learning, we aim to build a relationship between the source-domain outputs (*i.e.*, region proposals p_s and object defections d_s) and annotations $\hat{y}_s$. Given a source-domain image x_s and the network G, the supervised detection loss is defined as:

$$\mathcal{L}_{det}(x_s, \hat{y}_s; F) = \mathcal{L}_{rpn}(p_s, \hat{y}_s) + \mathcal{L}_{rcnn}(d_s, \hat{y}_s), \qquad (2)$$

where p_s and d_s denote the region proposals and object detection generated from G on source domain, $\mathcal{L}_{rpn}$ and $\mathcal{L}_{rcnn}$ represent RPN loss term and RCNN loss term respectively, as defined in [9].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3082687, IEEE Transactions on Multimedia

*2) **Image-level uncertainty-aware adversarial learning**:* For first-stage image-level domain adaptation, we propose an uncertainty-aware adversarial learning method to align the image-level features in source and target domains. Given image-level source and target features (f_{img_s} and f_{img_t}), region proposals (p_s and p_t), and the image-level domain classifier (C_{img}), the image-level uncertainty-aware adversarial loss is defined as:

$$\begin{aligned}\mathcal{L}_{img}^{ua}(C_{img}) = &\sum_{u,v}[\mathcal{E}_p(p_s^{(r,u,v)})\mathcal{L}_{img}^{ce}(o_{img_s}^{(u,v)}, l_s)+ \\ &\mathcal{E}_p(p_t^{(r,u,v)})\mathcal{L}_{img}^{ce}(o_{img_t}^{(u,v)}, l_t)],\end{aligned} \tag{3}$$

where $\mathcal{L}_{img}^{ce}$ denotes an image-level cross-entropy loss, $\mathcal{E}_p$ represents a proposal entropy function, $p^{(r,u,v)}$ is the r-th classification output located at (u, v) of region proposal prediction maps, r denotes the index of proposals with different scales and ratios in the same location, $o_{img_s}^{(u,v)}$ and $o_{img_t}^{(u,v)}$ are the activation located at (u, v) of the domain prediction maps generated from C_{img} in each domain, $l_s^{(u,v)}$ and $l_t^{(u,v)}$ are the pixel-level domain label locate at (u, v) of image-level domain labels (*i.e.*, 0 for source domain and 1 for target domain). The image-level cross-entropy loss $\mathcal{L}_{img}^{ce}$ in Eq. 3 is defined as:

$$\begin{aligned}\mathcal{L}_{img}^{ce}(o^{(u,v)}, l^{(u,v)}) = &- l^{(u,v)}\log(o^{(u,v)})- \\ &(1 - l^{(u,v)})\log(1 - o^{(u,v)}),\end{aligned} \tag{4}$$

For estimating the prediction confidence, we choose the lowest entropy of proposals in each location. Following information theory [49], the proposal entropy function $\mathcal{E}_p$ in Eq. 3 is defined as:

$$\begin{aligned}\mathcal{E}_p(p^{(r,u,v)}) = &\min_r[-p^{(r,u,v)} \cdot \log(p^{(r,u,v)}) \\ &- (1 - p^{(r,u,v)})(1 - \log(p^{(r,u,v)}))].\end{aligned} \tag{5}$$

*3) **Instance-level uncertainty-guided curriculum learning**:* For second-stage instance-level domain adaptation, we introduce an uncertainty-guided curriculum learning approach for more robust domain alignment. Given instance-level source-domain and target-domain features (f_{ins_s} and f_{ins_t}), detection results (d_s and d_t), instance-level proposal entropy (e_{ins_s} and e_{ins_t}) generated from proposal entropy using an ROI layer, and the instance-level domain classifier C_{ins}, the instance-level uncertainty-guided curriculum loss is defined as:

$$\begin{aligned}\mathcal{L}_{ins}^{ug}(C_{ins}) = \sum_k[&\mathcal{G}(\mathcal{E}_d(d_s^{(c,k)}), e_{ins_s}^{(k)}))\mathcal{L}_{ins}^{ce}(o_{ins_s}^{(k)}, l_s^{(k)})+ \\ &\mathcal{G}(\mathcal{E}_d(d_t^{(c,k)}), e_{ins_t}^{(k)}))\mathcal{L}_{ins}^{ce}(o_{ins_t}^{(k)}, l_t^{(k)})],\end{aligned} \tag{6}$$

where $\mathcal{L}_{ins}^{ce}$ denotes an instance-level cross-entropy loss, $\mathcal{G}$ represents a gate function, $\mathcal{E}_d$ represents a detection entropy function, $d^{(c,k)}$ represents the predicted probability of c-th class in k-th detection, $e^{(k)}$ is instance-level proposal entropy corresponding to k-th detection, $o_{ins_s}^{(k)}$ and $o_{ins_t}^{(k)}$ are the activation located at (k) of instance-level domain prediction vector generated from C_{ins} in each domain, $l^{(k)}$ is the instance-level domain label (*i.e.*, 0 for source domain and 1 for target domain) for k-th detection.

The instance-level cross-entropy loss $\mathcal{L}_{ins}^{ce}$ is defined by:

$$\begin{aligned}\mathcal{L}_{ins}^{ce}(o^{(k)}, l^{(k)}) = &- l^{(k)}\log(o^{(k)})- \\ &(1 - l^{(k)})\log(1 - o^{(k)})).\end{aligned} \tag{7}$$

The gate function $\mathcal{G}$ in Eq. 6 is defined as:

$$\mathcal{G}(\mathcal{E}_d(d^{(c,k)}), e_{ins}^k) = \begin{cases}\mathcal{E}_d(d^{(c,k)}) & e_{ins}^k < \xi \\ 0 & others\end{cases} \tag{8}$$

where ξ is a hyper-parameter to estimate whether image-level features are well-aligned or not. The instance-level adversarial learning can only be activated after image-level features are well-aligned, *i.e.*, the prediction entropy of proposals is low.

Following information theory [49], the detection entropy function $\mathcal{E}_d$ in Eq. 6 is defined as:

$$\mathcal{E}_d(d^{(c,k)}) = -\sum_c d^{(c,k)} \cdot \log d^{(c,k)}. \tag{9}$$

*4) **Optimization.**:* The overall loss of UaDAN is:

$$\mathcal{L}_{UaDAN} = \mathcal{L}_{det}(F) + \mathcal{L}_{img}^{ua}(C_{img}) + \mathcal{L}_{ins}^{ug}(C_{ins}). \tag{10}$$

The training objective of UaDAN is:

$$F^\star, C_{img}^\star, C_{ins}^\star = \arg\min_F \max_{C_{img}} \max_{C_{ins}} \mathcal{L}_{UaDAN} \tag{11}$$

We solve Eq. 11 by simultaneously optimizing G, C_{img} and C_{ins} until $\mathcal{L}_{UaDAN}$ converges.

D. Analysis

The major differences between our uncertainty-aware domain adaptation and traditional adversarial learning lie with uncertainty-aware adversarial learning and uncertainty-guided curriculum learning. We focus on instance-level domain adaptation to discuss the differences in this subsection.

Given instance-level source and target features (f_{ins_s} and f_{ins_t}) and the instance-level domain classifier C_{ins}, the instance-level traditional adversarial loss is defined as:

$$\mathcal{L}_{ins}(C_{ins}) = \sum_k[\mathcal{L}_{ins}^{ce}(o_{ins_s}^{(k)}, l_s^{(k)}) + \mathcal{L}_{ins}^{ce}(o_{ins_t}^{(k)}, l_t^{(k)})], \tag{12}$$

and the instance-level uncertainty-aware adversarial learning loss without curriculum learning is defined as:

$$\begin{aligned}\mathcal{L}_{ins}^{ua}(C_{ins}) = \sum_k[&\mathcal{E}_d(d_s)\mathcal{L}_{ins}^{ce}(o_{ins_s}^{(k)}, l_s^{(k)})+ \\ &\mathcal{E}_d(d_t)\mathcal{L}_{ins}^{ce}(o_{ins_t}^{(k)}, l_t^{(k)})].\end{aligned} \tag{13}$$

The differences between $\mathcal{L}_{ins}$ in Eq. 12 and $\mathcal{L}_{ins}^{ua}$ in Eq. 13 is that $\mathcal{L}_{ins}$ has the same weight while $\mathcal{L}_{ins}^{ua}$'s weight is decided by the prediction entropy. As studied in [69]–[71], well-aligned features usually produce confident predictions with low-entropy while under-aligned features often produce unconfident predictions with high-entropy. Utilizing entropy to adjust the loss weights can naturally protect well-aligned features from large re-alignment and focus more on aligning under-aligned features. Our uncertainty-aware adversarial learning is thus able to decrease the influence of domain discrepancy minimization on well-aligned features and lead to stable cross-domain alignment in each subtask.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3082687, IEEE Transactions on Multimedia

We provide certain theoretical analysis with a negative transfer scenario that often happens in unsupervised domain adaptation [72], [73] when the source-domain knowledge hurts the target-domain performance. Unsupervised domain adaptation aims to improve a predictive function G over unlabeled target domain D_t by learning transferable knowledge in labeled source domain D_s. We use $P_s(X,Y)$ and $P_t(X,Y)$ to denote the joint distribution in D_s and D_t, respectively, where X denotes input data and Y denotes labels. The objective of the traditional adversarial learning assumes that for any $x_t \in X_t$, there exists $x_s \in X_s$ such that $P_t(x_t, y_t) = P_s(x_s, y_s)$ [74]. However, $P_t(X,Y)$ and $P_s(X,Y)$ naturally have discrepancy between them. Considering the case with a well-classified (semantically well-aligned) target sample $x'_t \in X_t$ with domain specific features such that $P_t(x'_t, y'_t) \neq P_s(x_s, y_t)$ for any $x_s \in X_s$. If x'_t is trained with the objective of traditional adversarial learning that assumes $P_t(x_t, y_t) = P_s(x_s, y_s)$, it will be misaligned to incorrect classes due to negative transfer [73], leading to $P_t(x'_t, y'_t) \notin P_t(X,Y)$. Our uncertainty-aware adversarial learning can instead mitigate such negative transfer by adjusting the loss weight of the well-aligned sample x'_t (far from classifier boundary with entropy close to zero) using its entropy so that the well-aligned x'_t will not be further aligned. This largely helps to ensure $P_t(x'_t, y'_t) \in P_t(X,Y)$ and mitigate negative transfer effectively.

Compared with $\mathcal{L}_{ins}$ and $\mathcal{L}^{ua}_{ins}$, uncertainty-guided curriculum loss $\mathcal{L}^{ug}_{ins}$ in Eq. 6 further introduces a gate function $\mathcal{G}$ to activate instance-level alignment when the corresponding image-level representations are well-aligned (*i.e.*, low entropy). At early phase of training, source and target predictions are inaccurate and source knowledge is steadily transferred to target domain with the supervision of source ground-truth. Intuitively, aligning nonsensical instances could easily impair semantic structures and make negative impact. Harder subtask can make positive impact only when the easier subtask becomes relatively stable. Our developed uncertainty-guided curriculum learning solves the image-level alignment (easier subtask) and instance-level alignment (harder subtask) progressively to ensure more stable cross-domain alignment.

IV. Experiments

A. Experimental Setup

We follow the widely adopted protocol in domain adaptive object detection. Each task involves two datasets including a source dataset and a target dataset for training and evaluation. The training data consist of the labelled source training set and the unlabelled target training set. The validation set of the target datasets is used for evaluations in all methods. For the source and target datasets, only data with shared object categories are used in training and evaluations.

1) ***Dataset****:* Our experiments involves six public datasets including Mapillary Vistas [29], Cityscapes [27], Foggy Cityscapes [75], PASCAL VOC [25], Clipart [76], SIM10k [77]. More details of the six datasets are listed below.

• Mapillary Vistas [29] is a large-scale autonomous driving dataset with images recorded by different acquisition sensors. It contains 18,000 training images and 2,000 validation images, and the image resolution varies from 768×1024 to 4000×6000. Mapillary Vistas consists of 37 object categories.

• Cityscapes [27] is a widely used autonomous driving dataset with images captured with a unique vehicle mounted image acquisition system. It consists of 2,975 training images and 500 validation images with dense detection annotations, which are transformed from instance segmentation labels with 8 categories. All the images have the same resolution of 1024×2048 with common weather conditions.

• Foggy Cityscapes [75] dataset is established by applying fog simulation on the Cityscapes images [27]. The synthetic foggy images are rendered with visible images with common weather conditions and its depth counterpart in Cityscapes. The annotations are inherited from labels in Cityscapes.

• PASCAL VOC [25] is a large-scale real world dataset with two sub-datasets, *i.e.*, PASCAL VOC 2007 [78] and PASCAL VOC 2012 [79]. PASCAL VOC 2007 contains 2,501 for training and 2,510 for validation, while PASCAL VOC 2012 consists of 5,717 for training and 5,823 for validation. Bounding box annotations with 20 classes are provided.

• Clipart1k [76] dataset contains 1,000 comical images, in which 800 for training and 200 for validation. The manually created comical images have much dissimilarity as compared with real world images. It provides bounding box annotations with the same 20 categories as PASCAL VOC [25].

• SIM10k [77] dataset contains 10,000 synthetic images with automatically generated labels from computer games. All the images have the same resolution of 1052×1914. It provides bounding box annotations for cars.

2) ***Implementation details****:* Following [39], [40], [46], we adopt Faster R-CNN [9] as our object detection network. The backbone is initialized with deep convolutional layers of ResNet-50 [80], which is pre-trained on ImageNet [81]. The detection modules of Faster R-CNN (*i.e.*, RPN and RCNN) and the domain classifiers (*i.e.*, image-level and instance-level) are randomly initialized from a zero-mean Gaussian distribution with a standard deviation 0.01. During training, we use back-propagation and stochastic gradient descent (SGD) to optimize all the networks with a momentum of 0.9 and a weight decay of $5e-4$. The initial learning rate is set at 0.001 for $50k$ iterations and then reduced to 0.0001 for another $20k$ iterations.UaDAN uses one source image and one target image in each iteration as in [39]. For all experiments, the hyper-parameter ξ is set as 0.5. All experiments are implemented on one GPU by employing PyTorch toolbox, where the maximum memory usage is less than 9 GB. For evaluation, average precision (AP) of each category and mean average precision (mAP) of all categories are computed with an intersection over union (IoU) threshold 0.5.

B. Experimental Results

We evaluate our uncertainty-aware domain adaptation method in four different domain shift scenarios: 1) *Cross camera adaptation* with source and target images captured with different acquisition systems; 2) *Weather adaptation* with source images captured in good weather conditions and target

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3082687, IEEE Transactions on Multimedia

TABLE II
QUANTITATIVE COMPARISON OF UADAN WITH STATE-OF-THE-ART DOMAIN ADAPTIVE OBJECT DETECTION METHODS OVER THE CROSS CAMERA ADAPTATION TASK CITYSCAPES → MAPILLARY VISTAS: AP (%) OF EACH CATEGORY AND MAP (%) OF ALL CLASSES ARE EVALUATED OVER THE MAPILLARY VISTAS VALIDATION SET.

Methods	person	rider	car	truck	bus	train	motorbike	bicycle	mAP
Source only	31.3	30.0	48.9	20.3	23.1	7.1	21.2	24.3	25.8
DA [39]	34.2	28.8	55.3	20.0	19.5	17.8	23.9	28.2	28.4
SWDA [40]	33.4	29.3	50.9	23.4	26.6	23.8	28.2	25.0	30.1
CRDA [46]	34.0	**32.6**	51.4	23.3	24.0	22.4	28.2	27.0	30.4
CFA [45]	34.4	30.0	54.6	22.7	24.9	21.9	26.7	26.7	30.6
GPA [54]	35.9	31.1	**55.9**	20.1	23.1	25.6	28.2	28.6	31.0
UaDAN (Ours)	**36.1**	31.1	55.7	**24.3**	**27.0**	**28.3**	**30.1**	**28.9**	**32.7**

TABLE III
QUANTITATIVE COMPARISON OF UADAN WITH STATE-OF-THE-ART DOMAIN ADAPTIVE OBJECT DETECTION METHODS OVER THE WEATHER ADAPTATION TASK CITYSCAPES → FOGGY CITYSCAPES: AP (%) OF EACH CATEGORY AND MAP (%) OF ALL CLASSES ARE EVALUATED OVER THE FOGGY CITYSCAPES VALIDATION SET.

Methods	person	rider	car	truck	bus	train	motorbike	bicycle	mAP
Source only	26.9	26.9	38.2	18.3	32.4	9.6	25.8	28.6	26.9
DA [39]	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0
SWDA [40]	31.8	44.3	48.9	21.0	43.8	28.0	28.9	35.8	35.3
CRDA [46]	32.2	45.2	50.0	**30.3**	48.1	36.3	28.4	36.8	38.4
CFA [45]	**37.4**	45.3	53.5	25.8	**50.5**	31.3	30.2	**39.2**	39.1
GPA [54]	32.9	**46.7**	**54.1**	24.7	45.7	41.1	**32.4**	38.7	39.5
UaDAN (Ours)	36.5	46.1	53.6	28.9	49.4	**42.7**	32.3	38.9	**41.1**

TABLE IV
QUANTITATIVE COMPARISON OF UADAN WITH STATE-OF-THE-ART DOMAIN ADAPTIVE OBJECT DETECTION METHODS OVER THE DISSIMILAR ADAPTATION TASK PASCAL VOC → CLIPART1K: AP (%) OF EACH CATEGORY AND MAP (%) OF ALL CLASSES ARE EVALUATED OVER THE CLIPART1K VALIDATION SET. NOTE THAT AERO, BICY, BOTT, MOTOR, PERS AND TV ARE ABBREVIATIONS OF AEROPLANE, BICYCLE, BOTTLE, MOTORBIKE, PERSON AND TELEVISION, RESPECTIVELY.

Methods	aero	bicy	bird	boat	bott	bus	car	cat	chair	cow	table	dog	horse	motor	pers	plant	sheep	sofa	train	tv	mAP
Source only	4.3	75.0	31.3	21.5	4.8	64.5	23.2	**6.7**	33.1	7.1	32.5	4.9	56.5	66.7	49.6	51.9	9.4	20.5	34.5	21.6	31.0
DA [39]	17.9	**75.4**	32.9	23.7	10.9	35.8	31.0	3.8	42.2	70.8	13.6	16.9	31.3	86.7	61.7	53.9	10.0	25.2	29.9	18.7	34.6
SWDA [40]	6.8	50.2	33.0	19.2	13.3	61.4	35.4	5.9	**40.8**	54.5	**40.8**	24.6	53.9	76.7	63.6	61.0	10.0	**34.2**	32.1	19.9	36.8
CFA [45]	30.5	54.5	37.8	**27.7**	11.5	65.5	44.7	0.9	35.4	51.0	32.6	24.8	35.1	63.6	64.4	57.9	12.1	18.4	46.2	26.9	37.1
CRDA [46]	7.7	50.5	33.2	19.3	13.0	68.2	40.0	3.0	40.4	68.9	29.9	20.7	**57.2**	**86.9**	**68.6**	55.5	**15.1**	27.6	14.8	**30.5**	37.5
GPA [54]	12.2	54.5	40.3	25.6	16.8	**71.3**	39.9	4.2	38.9	**73.1**	21.6	25.7	54.5	63.6	63.1	58.6	13.6	11.1	44.7	31.6	38.3
UaDAN (Ours)	**35.0**	72.7	**41.0**	24.4	**21.3**	69.8	**53.5**	2.3	34.2	61.2	31.0	**29.5**	47.9	63.6	62.2	**61.3**	13.9	7.6	**48.6**	23.9	**40.2**

TABLE V
QUANTITATIVE COMPARISON OF UADAN WITH STATE-OF-THE-ART DOMAIN ADAPTIVE OBJECT DETECTION METHODS OVER THE SYNTHETIC-TO-REALISTIC ADAPTATION TASK SIM10K → CITYSCAPES: AP (%) OF CAR CLASS IS EVALUATED ON THE CITYSCAPES VALIDATION SET.

Methods	car AP
Source only	34.6
DA [39]	41.9
SWDA [40]	44.6
CFA [45]	46.0
CRDA [46]	46.6
GPA [54]	47.6
UaDAN (Ours)	**48.6**

images in foggy weather; 3) *Dissimilar adaptation* with source domain consisting of real-world images while target domain consisting of manually created images; and 4) *Synthetic-to-realistic adaptation* with source and target domains consisting of synthetic and real-world images. In each domain shift scenario, we compare UaDAN with a number of state-of-the-art unsupervised domain adaptive methods DA [39], SWDA [40], CRDA [46], CFA [45] and GPA [54].

*1) **Cross camera adaptation**:* Domain variance exists widely among images of different resolutions and qualities that are captured by using different acquisition sensors. For the task of object detection, we're facing more dramatic variations of object appearance in scale, viewpoints, etc. In this experiment, we study the effectiveness of our uncertainty-aware domain adaptation method while handling domain shifts among different real-image datasets. Specifically, we use 8 common object classes between the source dataset Cityscapes [27] and the target dataset Mapillary Vistas [29]. The validation set of the Mapillary Vista is used in evaluations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3082687, IEEE Transactions on Multimedia

(a) Cityscapes → Mapillary Vistas

(b) Cityscapes → Foggy Cityscapes

(c) PASCAL VOC → Clipart1k

(d) Sim10k → Cityscapes

Fig. 3. Qualitative comparison of UaDAN with 'Source only' (no adaptation) and GPA [54] over four domain adaptive detection tasks (a) Cityscapes → Mapillary Vistas, (b) Cityscapes → Foggy Cityscapes, (c) PASCAL VOC → Clipart1k and (d) Sim10k → Cityscapes: UaDAN outperforms GPA consistently by detecting more true positives (highlighted by yellow arrows) and less false positives (highlighted by red arrows) across all sample images. Bounding boxes of different color represent detection of different categories and a score threshold of 0.5 is used in visualization. Best viewed in color.

Table II shows comparisons of our UaDAN with state-of-the-art domain adaptive detection methods. We can see that UaDAN achieves the best detection with a mAP of 32.7%. For the classes (*e.g.*, train) which can hardly be detected by 'Source only' (with only 7.1% AP), UaDAN outperforms other methods by large margins (over 2.7% AP). A possible reason is that the 'train' features in target domain tend to be under-aligned due to large domain bias, and UaDAN can focus on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3082687, IEEE Transactions on Multimedia

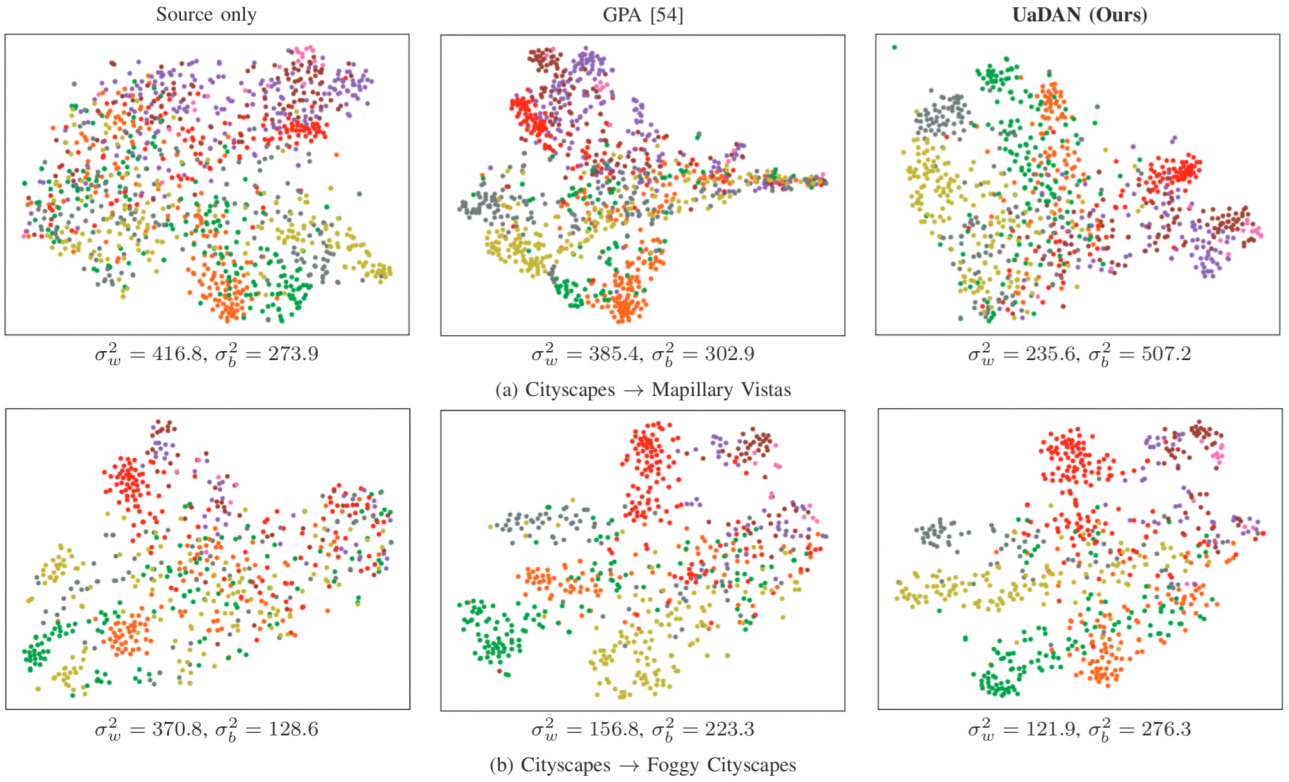

Fig. 4. Visualization of target-domain feature distributions with t-SNE [82]: We calculate within-class variance σ_w^2 and between-class variance σ_b^2 [83] for two domain adaptive object detection tasks Cityscapes → Mapillary Vistas in (a) and Cityscapes → Foggy Cityscapes in (b). It can be seen that our method outperforms 'Source only' (no adaptation) and state-of-the-art GPA [54] clearly. Note different colors represent different classes and best viewed in color.

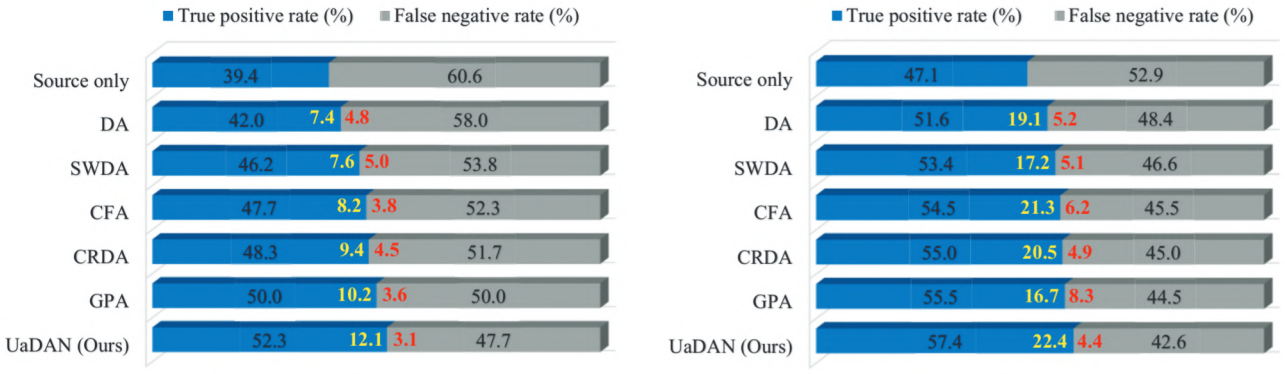

Fig. 5. Error analysis of UaDAN and state-of-the-art domain adaptive detection methods over two domain adaptive detection tasks: (a) Cityscapes → Mapillary Vistas and (b) PASCAL VOC → Clipart1k: UaDAN can generate most true positives which are falsely predicted by 'Source only' (highlighted in yellow color), and fewest false negatives which are accurately predicted by 'Source only' (highlighted in red color). Best viewed in color.

aligning under-aligned features while keeping the well-aligned features less affected.

*2) **Weather adaptation**:* Object detectors trained with normal-weather images usually do not perform well for images captured under adverse weather (*e.g.*, foggy, rainy and night-time). In practice, it's desired that object detectors can work well under different weather instead. In this study, we examine this issue by studying domain adaptation from normal weather to foggy weather. We use Cityscapes [27] as the source dataset, Foggy Cityscapes [75] as the target dataset, and the validation set of Foggy Cityscapes in evaluations.

Table III shows the comparison of UaDAN with state-of-the-art domain adaptive methods over Cityscapes → Foggy Cityscapes. We can see that UaDAN achieves the best detection accuracy with 41.1% mAP across all classes. Similarly for object classes (*e.g.*, train) which cannot be accurately detected using 'Source only' model (with only 9.6% AP), UaDAN outperforms other methods by over 1.6% in AP which is consistent with the experimental results on cross camera task. This experiment further verifies that our UaDAN can focus on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3082687, IEEE Transactions on Multimedia

aligning under-aligned samples.

*3) **Dissimilar adaptation**:* Both cross camera adaptation and weather adaptation work in certain similar domains. We perform one more experiment to study how UaDAN adapts across dissimilar domains from real to artistic images. We use PASCAL VOC [25] and Clipart [76] as source and target datasets which share 20 common object classes. The validation set of Clipart is used in evaluations.

Table IV shows the comparison of UaDAN with state-of-the-art methods on this new task. We can see that UaDAN achieves the best accuracy with 40.2% mAP across all classes. For object classes that cannot be detected well by 'Source only' (*e.g.*, aero, bottle, dog), UaDAN outperforms other methods by large margins (over 3.8% in AP for each category). This further shows that our uncertainty-aware adversarial learning can generalize to different domain adaption tasks.

*4) **Synthetic-to-realistic adaptation**:* Using synthetic images in deep network training has been attracting increasing interest in recent years. However, models trained using synthetic images usually experiences clear performance drops while applied to real images. In this experiment, we study how UaDAN performs for adaptation from synthetic to real images. We use SIM10k [77] and Cityscapes [27] as the source and target datasets that share only one object category 'car'. The validation set of Cityscapes is used in evaluations.

Table V shows the comparison of UaDAN with the state-of-the-art over the synthetic-to-real task. We can see that UaDAN achieves the best accuracy with a mAP 48.6%, showing that UaDAN is rather powerful in cross-domain detection task when the object category is small. We conjecture that the well-aligned car features could be over-aligned by the brute-force alignment while UaDAN keeps them less affected.

*5) **Qualitative comparison**:* The qualitative experimental results are well aligned with the quantitative results as shown in Fig. 3. We can see that UaDAN identifies more correct objects with less false positives than GPA [54] across all four target datasets. This also shows that UaDAN can keep well-aligned features less affected while brute-force alignment in GPA [54] maps them to incorrect categories. Detection may fail if 'Source only' predicts false positives with high confidence (low entropy) as shown in the first rows of Figs. 3 (b) and (c). Though UaDAN may generate false positives by keeping falsely identified well-aligned features from alignment, it can globally generate less false positives and more true positives by focusing on aligning poorly-aligned features in these scenarios.

C. Discussion

*1) **Feature Visualization**:* In the previous sections, we have shown that UaDAN achieves superior object detection performance as compared with state-of-the-art domain adaptive methods. To further analyse the behaviour of detection models, we employ t-SNE [82] to visualize the target-domain feature distributions as learnt by different cross-domain detection methods. In the visualization experiment, we randomly select 200 instances for each category (all instances are selected for categories with less than 200 instances). Fig. 4 shows the visualization, where the within-class variance and between-class variances are calculated for quantitative analysis. As Fig. 4 shows, the within-class and between-class variances are highly consistent with the object detection in Fig. 3.

*2) **Error Analysis**:* To further validate the effectiveness of the uncertainty awareness in protecting the well-aligned features from misalignment, we analyse the errors induced by domain adaptive methods as compared with 'Source only' (no adaptation). As Fig. 5 shows, we calculate the rate of true positives that are falsely predicted by 'Source only' (highlighted in yellow), and the rate of false negatives that are accurately predicted by 'Source only' (highlighted in red).

We can see that certain samples are accurately predicted by 'Source only' but falsely predicted by domain adaptive methods due to misalignment. UaDAN can alleviate this problem and produce less false negatives than other domain adaptive methods by keeping well-aligned features less affected. For samples that are falsely predicted by 'Source only', UaDAN can also produce more true positives than other methods by focusing on aligning under-aligned features.

*3) **Computational Overhead**:* We would clarify that UaDAN introduces very limited extra computation overhead in training (less than 0.0003 second per iteration on one GPU which translates to less than 0.1% in percentages) as compared to the traditional adversarial method [39], as entropy computation is simple and efficient. Meanwhile, UaDAN introduces no computation overhead during inference, as entropy computation is included in training stage only. However, it outperforms traditional adversarial methods by large margins, e.g. it outperform [39] by over 4.3% in AP over all domain adaptive detection tasks as shown in Tables II$\sim$V.

Since the framework of UaDAN is built upon Faster R-CNN, we further discuss the gap between them in in terms of accuracy, detection speed and model complexity. Faster R-CNN [9] with no adaptation trains a 'Source only' model by using labelled source data. The trained model does not perform well for target data due to domain gaps. The proposed UaDAN instead introduces uncertainty-aware domain classifiers that align source and target domains for better performance over target samples. Specifically, UaDAN achieves much better AP (improved by over 6.9%) over all domain adaptive detection tasks as shown in Tables II$\sim$V. In addition, it has the same detection speed and model complexity as Faster R-CNN since domain adaptation modules are included in training stage only.

*4) **Training and Testing Processing**:* We also compute the training and testing losses of the UaDAN in every 10^3 iterations. As shown in Fig. 6, both training and testing losses decrease quickly during the first 10k iterations and then fluctuate in a small range after 50k iterations.

D. Ablation Studies

We perform a series of ablation experiments to study how UaDAN components contribute to overall performance. Seven models are trained as listed in Table VI: 1) *Baseline* which is 'Source only' as trained by source samples without adaptation; 2) *Image-level AL* which is image-level adversarial model as trained by $\mathcal{L}_{det}$ and image-level adversarial loss $\mathcal{L}_{img}$;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3082687, IEEE Transactions on Multimedia

TABLE VI
ABLATION STUDY OF OUR METHOD OVER DOMAIN ADAPTIVE DETECTION TASK CITYSCAPES → MAPILLARY VISTAS: UNCERTAINTY-AWARE ADVERSARIAL LEARNING (UAAL) OUTPERFORMS THE TRADITIONAL ADVERSARIAL LEARNING (AL) CONSISTENTLY AT BOTH IMAGE LEVEL AND INSTANCE LEVEL. UNCERTAINTY-GUIDED CURRICULUM LEARNING (UGCL) FURTHER BOOSTS THE PERFORMANCE OF THE ADVERSARIAL MODEL. NOTE THAT MAP (%) IS EVALUATED OVER THE MAPILLARY VISTAS VALIDATION SET.

Method	$\mathcal{L}_{det}$	$\mathcal{L}_{img}$	$\mathcal{L}_{img}^{ua}$	$\mathcal{L}_{ins}$	$\mathcal{L}_{ins}^{ua}$	$\mathcal{L}_{ins}^{ug}$	mAP
Baseline	✓						25.8
Image-level AL	✓	✓					29.7
Image-level UaAL	✓		✓				30.8
Instance-level AL	✓			✓			26.7
Instance-level UaAL	✓				✓		29.8
UaDAN w/o UgCL	✓		✓		✓		31.5
UaDAN	✓		✓		✓	✓	**32.7**

Fig. 6. The training and testing losses of our method over domain adaptive detection task Cityscapes → Mapillary Vistas.

3) *Image-level UaAL* which is image-level uncertainty-aware adversarial model as trained by $\mathcal{L}_{det}$ and entropy-weighed image-level adversarial loss $\mathcal{L}_{img}^{ua}$; 4) *Instance-level AL* which is instance-level adversarial model as trained by $\mathcal{L}_{det}$ and instance-level adversarial loss $\mathcal{L}_{ins}$; 5) *Instance-level UaAL* which is instance-level entropy-weighed adversarial model as trained by $\mathcal{L}_{det}$ and instance-level entropy-weighed adversarial loss $\mathcal{L}_{ins}^{ua}$; 6) *UaDAN w/o UgCL* which is UaDAN model without uncertainty-guided curriculum learning. (trained by $\mathcal{L}_{det}$, $\mathcal{L}_{img}^{ua}$ and $\mathcal{L}_{ins}^{ua}$); and 7) *UaDAN* which is the complete UaDAN model as trained by $\mathcal{L}_{det}$, $\mathcal{L}_{img}^{ua}$, $\mathcal{L}_{ins}^{ua}$, and $\mathcal{L}_{ins}^{ug}$.

As Table VI shows, *Image-level AL* and *Instance-level AL* both outperform the *Baseline* consistently, which demonstrates the importance of feature representation alignment at both image level and instance level in cross-domain detection tasks. Specifically, the gain of *Image-level AL* is much higher than *Instance-level AL*, largely because harder instance-level alignment can work only when the easier image-level alignment works. In addition, we can observe that *Image-level UaAL* and *Instance-level UaAL* outperform *Image-level AL* and *Instance-level AL* consistently in both image-level and instance-level detection tasks, demonstrating the importance of uncertainty-aware alignment in keeping well-aligned features less affected. Further, *UaDAN w/o UgCL* outperforms both *Instance-level UaAL* and *Instance-level UaAL*, which shows that Image-level UaAL and Instance-level UaAL are complementary. Finally, *UaDAN* outperforms *UaDAN w/o UgCL* with a large margin, which verifies the effectiveness of our proposed uncertainty-guided curriculum learning.

TABLE VII
THE SENSITIVITY OF PARAMETER ξ: UADAN OBTAINS THE BEST PERFORMANCE CONSISTENTLY WHEN $\xi = 0.5$. NOTE THAT MAP (%) IS EVALUATED OVER THE VALIDATION SET OF EACH DOMAIN ADAPTIVE DETECTION TASK.

ξ	0	0.25	0.5	0.75	1
Cityscapes → Mapillary Vistas	30.8	31.5	**32.7**	32.1	31.5
Cityscapes → Foggy Cityscapes	39.5	40.5	**41.1**	40.7	40.0
PASCAL VOC → Clipart1k	37.6	38.7	**40.2**	39.6	38.8

we also study the sensitivity of parameter ξ over three domain adaptive detection tasks as shown in Table VII. Specifically, *UaDAN* degrades to *Image-level UaAL* when $\xi = 0$, where the instance-level alignment is not activated as entropy cannot be negative. It degrades to *UaDAN w/o UgCL* (described in the Section of Ablation Studies) when $\xi = 1$, where the instance-level alignment is always activated as the entropy is smaller than 1. In addition, the best performance is obtained when $\xi = 0.5$ consistently over all three tasks.

V. CONCLUSION

This paper presents an uncertainty-aware domain adaption technique for unsupervised domain adaptation in object detection. We design an uncertainty-aware adversarial learning algorithm that can keep well-aligned features less affected in both proposal generation and object detection tasks. In addition, we design a uncertainty-guided curriculum learning algorithm that can alleviate the side effect of the traditional adversarial learning in handling harder detection tasks. Extensive experiments over four challenging cross-domain detection tasks demonstrate the effectiveness of the proposed method. Moving forwards, we will explore how to adapt the proposed technique to other domain adaptive tasks such as image classification and semantic segmentation.

ACKNOWLEDGMENT

This research was conducted at Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU),

which is a collaboration between Singapore Telecommunications Limited (Singtel) and Nanyang Technological University (NTU) that is funded by the Singapore Government through the Industry Alignment Fund - Industry Collaboration Projects Grant.

REFERENCES

[1] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *IEEE Computer Vision and Pattern Recognition*. IEEE, 1998, pp. 555–562.

[2] P. Viola, M. Jones *et al.*, "Robust real-time object detection," *International journal of computer vision*, vol. 4, no. 34-47, p. 4, 2001.

[3] J. C. Nascimento and J. S. Marques, "Performance evaluation of object detection algorithms for video surveillance," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 761–774, 2006.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[5] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1896–1909, 2013.

[6] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[8] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[12] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[14] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[15] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 9627–9636.

[16] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[18] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3588–3597.

[19] Z. Cai and N. Vasconcelos, "Cascade r-cnn: high quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1483–1498, 2021.

[20] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 821–830.

[21] T. Chen, S. Lu, and J. Fan, "S-cnn: Subcategory-aware convolutional networks for object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 10, pp. 2522–2528, 2017.

[22] H. Qiu, H. Li, Q. Wu, F. Meng, L. Xu, K. N. Ngan, and H. Shi, "Hierarchical context features embedding for object detection," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3039–3050, 2020.

[23] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 781–10 790.

[24] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[25] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[27] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[28] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[29] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4990–4999.

[30] S. D. Roy, T. Mei, W. Zeng, and S. Li, "Towards cross-domain learning for social video popularity prediction," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1255–1267, 2013.

[31] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 64–78, 2014.

[32] J. Huang, D. Guan, A. Xiao, and S. Lu, "Fsdr: Frequency space domain randomization for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[33] Y. Zhai, S. Lu, Q. Ye, X. Shan, J. Chen, R. Ji, and Y. Tian, "Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9021–9030.

[34] J. Huang, D. Guan, A. Xiao, and S. Lu, "Cross-view regularization for domain adaptive panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[35] H. Yan, Z. Li, Q. Wang, P. Li, Y. Xu, and W. Zuo, "Weighted and class-specific maximum mean discrepancy for unsupervised domain adaptation," *IEEE Transactions on Multimedia*, vol. 22, no. 9, pp. 2420–2433, 2019.

[36] J. Huang, S. Lu, D. Guan, and X. Zhang, "Contextual-relation consistent domain adaptation for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 705–722.

[37] D. Guan, J. Huang, S. Lu, and A. Xiao, "Scale variance minimization for unsupervised domain adaptation in image segmentation," *Pattern Recognition*, vol. 112, p. 107764, 2021.

[38] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[39] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3339–3348.

[40] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6956–6965.

[41] C. Zhang, Z. Li, J. Liu, P. Peng, Q. Ye, S. Lu, T. Huang, and Y. Tian, "Self-guided adaptation: Progressive representation alignment for domain adaptive object detection," *IEEE Transactions on Multimedia*, 2021.

[42] C. Zhuang, X. Han, W. Huang, and M. R. Scott, "ifan: Image-instance full alignment networks for adaptive object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[43] Z. He and L. Zhang, "Domain adaptive object detection via asymmetric tri-way faster-rcnn," in *European conference on computer vision*, 2020.

[44] G. Zhao, G. Li, R. Xu, and L. Lin, "Collaborative training between region proposal localization and classification for domain adaptive object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 86–102.

[45] Y. Zheng, D. Huang, S. Liu, and Y. Wang, "Cross-domain object detection through coarse-to-fine feature adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 766–13 775.

[46] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, "Exploring categorical regularization for domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 724–11 733.

[47] J. Zhang, J. Huang, Z. Luo, G. Zhang, and S. Lu, "Da-detr: Domain adaptive detection transformer by hybrid attention," *arXiv preprint arXiv:2103.17084*, 2021.

[48] F. Zhan, C. Xue, and S. Lu, "Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9105–9115.

[49] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[50] L. Nie, X. Wang, J. Zhang, X. He, H. Zhang, R. Hong, and Q. Tian, "Enhancing micro-video understanding by harnessing external sounds," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1192–1200.

[51] P. Jing, Y. Su, L. Nie, and H. Gu, "Predicting image memorability through adaptive transfer learning from external sources," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 1050–1062, 2016.

[52] M. S. M. Azzam, W. Wu, W. Cao, S. Wu, and H.-S. Wong, "Ktransgan: Variational inference-based knowledge transfer for unsupervised conditional generative learning," *IEEE Transactions on Multimedia*, 2020.

[53] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 457–11 466.

[54] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-domain detection via graph-induced prototype alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 355–12 364.

[55] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[56] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in neural information processing systems*, 2010, pp. 1189–1197.

[57] F. Khan, B. Mutlu, and J. Zhu, "How do humans teach: On curriculum learning and teaching dimension," in *Advances in neural information processing systems*, 2011, pp. 1449–1457.

[58] A. Pentina, V. Sharmanska, and C. H. Lampert, "Curriculum learning of multiple tasks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5492–5500.

[59] N. Sarafianos, T. Giannakopoulos, C. Nikou, and I. A. Kakadiaris, "Curriculum learning of visual attribute clusters for multi-task classification," *Pattern Recognition*, vol. 80, pp. 94–108, 2018.

[60] T. Matiisen, A. Oliver, T. Cohen, and J. Schulman, "Teacher-student curriculum learning," *IEEE transactions on neural networks and learning systems*, 2019.

[61] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5901–5910.

[62] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang, "Multi-modal curriculum learning for semi-supervised image classification," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3249–3260, 2016.

[63] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M. R. Scott, and D. Huang, "Curriculumnet: Weakly supervised learning from large-scale web images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 135–150.

[64] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," *International Journal of Computer Vision*, vol. 127, no. 4, pp. 363–380, 2019.

[65] X. Chang, Z. Ma, X. Wei, X. Hong, and Y. Gong, "Transductive semi-supervised metric learning for person re-identification," *Pattern Recognition*, p. 107569, 2020.

[66] Q. Yu, D. Ikami, G. Irie, and K. Aizawa, "Multi-task curriculum framework for open-set semi-supervised learning," in *European conference on computer vision*, 2020.

[67] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2020–2030.

[68] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, 2015, pp. 1180–1189.

[69] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Advances in neural information processing systems*, 2005, pp. 529–536.

[70] C.-H. Lee, S. Wang, F. Jiao, D. Schuurmans, and R. Greiner, "Learning to model spatial dependency: Semi-supervised discriminative random fields," in *Advances in Neural Information Processing Systems*, 2007, pp. 793–800.

[71] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.

[72] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[73] Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, "Characterizing and avoiding negative transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 293–11 302.

[74] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[75] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[76] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5001–5009.

[77] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?" in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 746–753.

[78] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge 2007 (voc2007) results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[79] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[80] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[81] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[82] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[83] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.