Domain-Adaptive Object Detection via Uncertainty-Aware Distribution Alignment

Dang-Khoa Nguyen
National Chiao Tung University
Hsinchu, Taiwan
basicskywards.eic07g@nctu.edu.tw

Wei-Lun Tseng
National Chiao Tung University
Hsinchu, Taiwan
erictseng.eed06g@nctu.edu.tw

Hong-Han Shuai*
National Chiao Tung University
Hsinchu, Taiwan
hhshuai@nctu.edu.tw

ABSTRACT

Domain adaptation aims to transfer knowledge from the source data with annotations to scarcely-labeled data in the target domain, which has attracted a lot of attention in recent years and facilitated many multimedia applications. Recent approaches have shown the effectiveness of using adversarial learning to reduce the distribution discrepancy between the source and target images by aligning distribution between source and target images at both image and instance levels. However, this remains challenging since two domains may have distinct background scenes and different objects. Moreover, complex combinations of objects and a variety of image styles deteriorate the unsupervised cross-domain distribution alignment. To address these challenges, in this paper, we design an end-to-end approach for unsupervised domain adaptation of object detector. Specifically, we propose a Multi-level Entropy Attention Alignment (MEAA) method that consists of two main components: (1) Local Uncertainty Attentional Alignment (LUAA) module to accelerate the model better perceiving structure-invariant objects of interest by utilizing information theory to measure the uncertainty of each local region via the entropy of the pixel-wise domain classifier and (2) Multi-level Uncertainty-Aware Context Alignment (MUCA) module to enrich domain-invariant information of relevant objects based on the entropy of multi-level domain classifiers. The proposed MEAA is evaluated in four domain-shift object detection scenarios. Experiment results demonstrate state-of-the-art performance on three challenging scenarios and competitive performance on one benchmark dataset.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**; *Computer vision tasks.*

KEYWORDS

Unsupervised domain adaptation; object detection; uncertainty; adversarial learning

*Corresponding author.

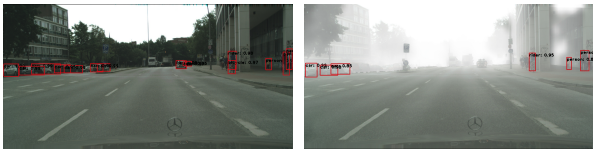
MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413553

ACM Reference Format:
Dang-Khoa Nguyen, Wei-Lun Tseng, and Hong-Han Shuai. 2020. Domain-Adaptive Object Detection via Uncertainty-Aware Distribution Alignment. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413553

1 INTRODUCTION

Deep neural networks have achieved high performance on object detection tasks thanks to learning feature representation from a large amount of training data with labels [5, 20, 21]. The generalizability of the deep neural networks, i.e., the ability of training on one domain images and performing well on another domain images, saves time and annotation costs, which accelerates the development of multimedia applications in real-world problems, especially when annotations for images are limited or difficult to acquire.



(a) Source Domain (Cityscapes) (b) Target Domain (Foggy Cityscapes)

Figure 1: Domain-adaptive object detection examples from Cityscapes (a) and Foggy Cityscapes dataset (b), respectively. The model is trained on the Cityscapes dataset and directly performs detection on the Foggy Cityscapes images. It clearly shows that the detection precision is considerably degraded due to the large domain gap between the two different domains, normal to foggy weather.

However, image distributions from two different domains usually exhibit a large domain gap, which leads to the degradation of detection performance in terms of accuracy and robustness. To address this challenge, a model trained on one domain (so-called *source domain*) is required to achieve generalization ability when conducting the inference on a new domain (so-called *target domain*), in which the image style, object appearance, background, *etc.* are considerably different from the source images. For instance, we train a domain-adaptive object detector on Cityscapes [4] images (all images taken in good weather condition) with bounding box annotations and Foggy Cityscapes [23] (all images represent foggy scenes) without any labels. After training, the trained model performs object detection on foggy images from the Foggy Cityscapes dataset. In this scenario, unsupervised domain-adaptive object detection is the most appropriate and efficient approach, which transfers the knowledge learned from label-rich source images to no-label target

images. Recently, many prior studies [2, 3, 22, 30, 33] have investigated domain-shift problems by aligning distribution at image and instance levels. However, image-level alignment might inaccurately transfer non-relevant backgrounds. Moreover, realizing hard-to-aligned features at instance level still remains challenging.

To address the above challenges, we aim to design an end-to-end unsupervised domain-adaptive object detector. Specifically, we propose a Multi-level Entropy Attention Alignment (MEAA) method which is composed of two components: (1) Local Uncertainty Attentional Alignment (LUAA) module to facilitate the model to better perceive structure-invariant objects by utilizing information theory to measure the uncertainty of each local region in a whole image and (2) Multi-level Uncertainty-aware Context Alignment (MUCA) module to enrich domain-invariant information of relevant objects at multiple levels of feature representation (*e.g.* local region, image, instance). For LUAA, we use the entropy of the outputs of pixel-wise domain discriminator to construct a local entropy-based feature map as an attention map, which enforces the model to pay more attention to complex objects. For example, the structures of objects are viewed differently from different angles or discontinuous patterns. It also diminishes the effect of unrelated objects and background (e.g., road, sky). As such, LUAA enhances the local features and makes the model detect objects better. On the other hand, MUCA aims to learn domain-invariant information of relevant objects. In addition, we hypothesize that objects of interest are more likely with high entropy due to its varying structure and texture. Therefore, the uncertainty-aware attentional method facilitates the model by semantically focusing on objects of interest to better realize domain-invariant representation in feature space.

The contributions of this work can be summarized as follows:

- A novel multi-level entropy attentional alignment framework is presented, which improves the performance of the unsupervised domain adaptation of the Faster-RCNN object detector.
- Two modules are designed to utilize multi-level domain-invariant features in terms of uncertainty perspective. The uncertainty-aware attentional alignment module aligns high-level features and low-level features by realizing structure-invariant regions of objects. The uncertainty-aware context alignment module enriches the model with uncertainty-weighted context vectors.
- Intensive experiments on various public benchmark domain-shift scenarios show that our framework achieves state-of-the-art performance on three challenging scenarios and competitive performance on one benchmark dataset.

2 RELATED WORK

2.1 Object Detection

Object detection is a classic problem in the computer vision area. In recent years, the outstanding performance of deep learning has shown superiority over other machine learning methods. The mainstream of deep learning approaches for object detection can be categorized into two parts: region proposal-based and region proposal-free. Region proposal-based methods are the well-known RCNN [10] family. First, regions of interest (RoI) are searched. With RoI as the proposal, after doing feature extraction of RoI, the predicted bounding boxes and corresponding category are produced. However, the whole inference time of [10] is too much. Fast-RCNN [9] is presented to improve processing time. In [9], RoI pooling layer is proposed to deal with the computational complexity of RoI feature extraction. Nevertheless, the searching time of RoI in [9] is still too much. The improved version of [9] is Faster-RCNN [21]. In [21], the region proposal network (RPN) is proposed to select RoI more efficiently. This makes total inference time approach to real-time, helping to deploy [21] on robotics or autonomous vehicles.

2.2 Domain Adaptation

Previous work in the unsupervised domain adaptation (UDA) setting are investigated in different topics including image classification [8, 18, 19, 31], semantic segmentation [1, 13, 23, 24, 26, 28] and object detection [2, 3, 14, 22, 30, 33]. Long *et al.* [19] study minimizing the maximum mean discrepancy between two domains. Ganin *et al.* [8] propose an adversarial loss with a domain classifier to learn domain-invariant and discriminative features. Li *et al.* [18] exploit an adversarial learning to simultaneously learn domain-wise and class-wise distribution between source and target domain. Wang *et al.* [29] perform dynamic distribution alignment for manifold domain adaptation by learning a domain-invariant classifier in Grassmann manifold with structural risk minimization. Zhuo *et al.* [34] propose an attention transfer schema for domain adaptation via minimizing the classification loss of source data with labels and the unsupervised correlation alignment loss. Besides leveraging adversarial learning, the synthetic dataset with full annotations is also exploited. Zhang *et al.* [32] propose a method aiming at leveraging the knowledge in the synthetic datasets to enhance pose estimator training on real-world datasets.

In the line of domain-adaptive object detection, Domain Adaptive Faster-RCNN [3] is the pioneering work that applies adversarial training losses [7, 27] into Faster-RCNN [9] in order to align domain-invariant features between source and target domain images. Specifically, its efforts to close distribution discrepancy at image and instance levels between two domains which are carried out at the backbone network and at the instance-level features of regions of interest (ROI) operation right before the final object classification and regression losses. Recently, Domain Adaptation Faster-RCNN series has achieved successful improvement [2, 22, 30, 33]. For instance, Saito *et al.* [22] conduct distribution alignment in feature space at a lower level of the backbone network and at a high level right before Region Proposal Network (RPN) which enhances image-level and instance-level feature alignment via adversarial training. Xu *et al.* [30] recently explore the classification ability and the categorical consistency between image-level and instance-level prediction which make the model deliberately align regions of relevant objects and hard aligned instances. Chen *et al.* [2] further investigate calibrates the transferability of feature representation at local-region, image, instance levels. Specifically, it utilizes CycleGAN to generate interpolated images to initially reduce global discriminability and propose a context-aware instance-level alignment to enhance the instance-level feature alignment by capturing global context information. Local feature masks are also used to provide semantic guidance for following global alignment. In this

approach, the model requires a computationally expensive step compared with other approaches. Furthermore, realizing hard-aligned regions of objects of interest still remains challenging to the prior approaches which motivate our method.

3 METHODS

In this section, we present our proposed method in detail. The essence of our method is to learn domain-invariant features by leveraging information theory to measure the uncertainty of objects and enrich the discriminative representation in the feature space. To this end, we propose a multi-level entropy attentional alignment (MEAA) method for integrating into an end-to-end adversarial domain adaptation framework which is illustrated in Fig. 2.

Specifically, in Fig. 2, the proposed MEAA framework is composed of the LUAA module and the MUCA module. During the training, a pair of input images (one randomly-selected source image and one target image) will be fed into an image encoder (F1). The output feature map from F1 will be fed into the lower path (LUAA) for enhancing the robustness of local features. For LUAA, we use the output of the pixel-wise domain discriminator (D1) to calculate the entropy, and then utilize this entropy to generate uncertainty-aware attentional feature maps which are subsequently fed into the upper path (MUCA) for enriching domain-invariant information of relevant objects. Both paths leverage Gradient Reverse Layers (GRLs) and domain discriminators to perform adversarial learning. In other words, the entropy map serves as the attention mechanism, which is important to domain adaptation. For MUCA, after obtaining the uncertainty-aware attentional feature maps from the output of pixel-wise domain discriminator (Eq. 8, 9 and 10), i.e., p4, we calculate the entropy map of p4 and fuse it with the output of feature extractor F4 to generate uncertainty-aware context vector h4 (Eq. 10). The following two layers of MUCA repeat the same operation to derive the multi-level features. In the last part, we will use these uncertainty-aware context vectors from multiple layers by concatenating them with the instance-level features at the end of Faster-RCNN backbone network, which provide new augmented instance-level features (Eq. 12).

3.1 Preliminaries and Motivations

In domain adaptive object detection, the model performs object classification and bounding box localization simultaneously. Formally, we denote source domain dataset $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ where for index i, $x_i^s \in \mathcal{R}^{H\times W\times 3}$ is the i-th image from source domain from N_s samples, and its bounding box annotation and the class label is $y_i^s \in \mathcal{R}^{k\times 5}$ which represents a list of the bounding box(es) coordinate and corresponding class(es). And the target domain $\mathcal{D}_t = \{x_j^t\}_{j=1}^{N_t}$ where assuming no bounding box annotations and object labels available. The data distributions between these two domains are different. Therefore, the ultimate goal is to design a novel generalized framework that enables Faster-RCNN learning domain-invariant features of objects from images with annotations and images without annotations during training to well perform on target data for testing.

Main challenges involve in cross-domain object detection including (1) domain gap caused by distribution discrepancy at not only global (*images*) but also local level (*objects*) that is more challenging than domain adaptation for classification task where a model only need to capture the global domain shift; (2) learning discriminative feature representation for both the multiple class-wise classifications and the bounding box regression simultaneously.

To resolve these challenges, our domain adaptive detection approach the problems from the following perspective. Specifically, feature distribution alignment is activated at multiple levels of the backbone network. In addition, information purity of input images is measured via adversarial learning which is fused into the feature representation to enforce the alignment between two domains effectively.

3.2 Uncertainty-Aware Attentional Alignment Module

We observe that each local region in a whole image contains different objects or parts of objects (*e.g.* road, sky, car, *etc.*) which represented by a certain amount of pixels with a different intensity which can be viewed as a degree of uncertainty. Therefore, some regions exhibit more important information than another in terms of feature discriminability and transferability. Moreover, the structure and texture are distinct among different instances and relatively similar if instances are in the same category. The Structure of images is shown as the most decisive factor for the unsupervised semantic segmentation task [1]. Inspired by [1], we hypothesize that the structural and textural contents of each region in images are also the decisive elements for learning domain-invariant features in the unsupervised domain-adaptive task. To semantically align distributions between source and target images, the model is required to identify domain-invariant representation in feature space. To this end, inspired by [2, 28], we facilitate the model to better perceive structure-invariant objects by utilizing information theory to measure the uncertainty of each local region in a whole image and then fuse them into conventional feature maps which are turned into new weighted maps, namely an uncertainty-aware feature maps.

In detail, an input image x_i is fed into the network. The feature maps $f_1^{H\times W}$ with size W width and H height is generated by $F_1(x_i)$ in Eq. (2) where F_1 is the first feature extractor. To learn domain-invariant feature at the local level, inspired by [22], D_1 is used as a pixel-wise domain discriminator. Therefore, the pixel-wise loss function $\mathcal{L}_1$ in adversarial training is defined as follows,

$$\begin{aligned}\mathcal{L}_1 = &\frac{1}{N_s HW}\sum_{i=1}^{N_s}\sum_{j=1}^{H\times W}\log(D_1(F_1(x_i^s)_j))^2 \\ &+\frac{1}{N_t HW}\sum_{i=1}^{N_t}\sum_{j=1}^{H\times W}\log(1-D_1(F_1(x_i^t)_j))^2,\end{aligned} \tag{1}$$

where N_s and N_t are the number of images from source and target, $(F_1(x_i^s)_j$ and $(F_1(x_i^t)_j$ denote the element j-th of the feature maps from the i-th source and target image input, respectively.

$$f_1^{H\times W} = F_1(x_i) \tag{2}$$

The pixel-wise probability output of discriminator D_1 is denoted as $d_1^{H\times W}$ in Eq. (3). We here compute pixel-level entropy map $E_1^{H\times W}$ as in Eq. (4) as following the Shannon Entropy. The entropy map

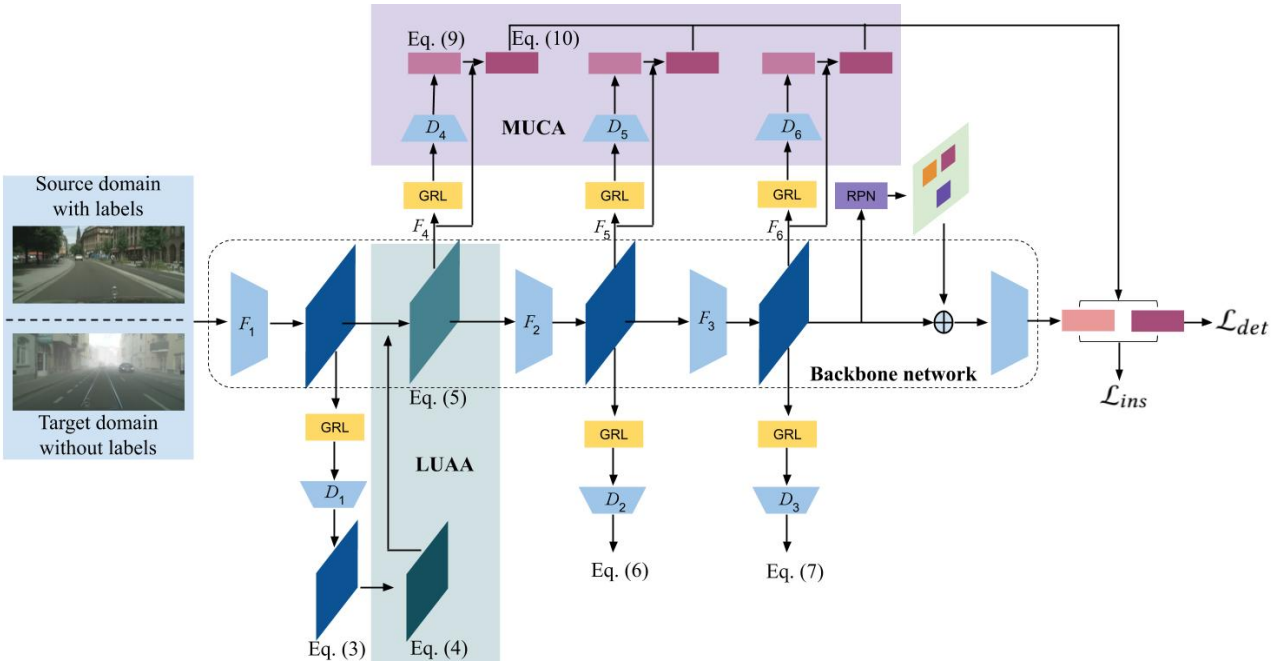


Figure 2: An overview of our MEAA framework composed of two proposed modules, a local uncertainty-aware attentional alignment (LUAA) module, and a multi-level uncertainty-aware context alignment (MUCA) module. Best view in color.

$E_1^{H\times W}$ is composed of pixel-wise entropies normalized to the range $[0, 1]$. Each element of entropy map represents a local region of an image corresponding to a receptive field thanks to the principle of convolutional operations. According to Shannon Entropy property, less uncertainty (low entropy) implies a higher purity of information and vice versa. With the hypothesis that in terms of structure and texture, the entropy for homogeneous regions of images is lower than that of the heterogeneous ones, for instance, entropies of road or sky objects are lower than other more complex ones such as a person, motorbike. Therefore, we directly fuse the entropy map into the conventional feature map extracted as in Eq. (2) which behaves as attention-like property in Eq. (5). Intuitively, this will assign lower weights on low entropy regions of images, even though they are higher purity of information or less uncertainty. As a result, high entropy regions (*e.g.* structure-complex objects including person, truck, motorbike, *etc.*) of images with higher uncertainty are assigned higher weights. This semantically enhances the uncertainty awareness toward objects of interest.

$$d_1^{H\times W} = D_1(f_1^{H\times W}) \tag{3}$$

$$E_1^{H\times W} = H(d_1^{H\times W}) = -d_1^{H\times W} \odot \log(d_1^{H\times W}) \tag{4}$$

where $H(\cdot)$ is the entropy function, "$\odot$" and "log" stands for Hadamard product and point-wise logarithm, respectively.

$$h_1 = f_1^{H\times W} \odot E_1^{H\times W} \tag{5}$$

The uncertainty-aware feature maps h_1 computed in Eq. (5) are then fed into the next feature extractor F_2. Inspired by [2, 22], to further enhance learning domain-invariant, another two domain discriminators D_2 defined in Eq. (6) and D_3 defined in Eq. (7) are deployed at middle and high levels of the network, respectively. The adversarial loss is expressed as follows,

$$\mathcal{L}_2 = \mathbb{E}[\log(D_2(F_2(h_1^s)))] + \mathbb{E}[\log(1 - D_2(F_2(h_1^t)))], \tag{6}$$

where h_1^s and h_1^t are the uncertainty-aware feature maps from source and target images, respectively,

$$\mathcal{L}_3 = \mathbb{E}[\log(D_3(F_3(f_2^s)))] + \mathbb{E}[\log(1 - D_3(F_3(f_2^t)))]. \tag{7}$$

To enable adversarial training for domain adaptation, following the design protocol by previous work [2, 3, 22], we apply the gradient reverse layer (GRL) [7] which its position is between the domain classifier and the feature-extracting network of the detector and it works as an identity function in forwarding pass and reverses the gradient during backpropagation.

3.3 Multi-level Uncertainty-Aware Context Alignment Module

We design a Multi-level Uncertainty-aware Context Alignment (MUCA) module to automatically capture domain-invariant structures among target-specific and source-specific feature representation by enriching domain-invariant information of relevant objects at multiple levels of the backbone network and at multi-level feature representation of objects including local regions, image, instance.

Specifically, during the training process, the weights of lower layers will be updated with respect to the gradients of adversarial and detection losses. This leads the network to learn domain-invariant

features at both high layers and low layers. However, these gradients may be weaker and weaker at lower layers which deteriorate awareness of feature uncertainty and degenerate discriminability and transferability of domain-invariant features. Therefore, inspired by [2, 22], we propose a multi-level uncertainty-aware context alignment. Technically, an auxiliary feature extractor F_4 is used to generate low-level context feature vectors denoted as f_4 which is then fed into a domain discriminator D_4 in Eq. (8) as below,

$$p_4 = D_4(F_4(h_1^{H\times W})). \tag{8}$$

Next, we measure the uncertainty of input images by computing entropy vectors E_4 based on the probability output of domain classifier D_4 as in Eq. (9) below,

$$E_4 = H(p_4) = -p_4 \odot \log(p_4). \tag{9}$$

To enable uncertainty-aware property, entropy vectors are fused into f_4 feature vectors, resulting in uncertainty-aware context vectors denoted as h_4 in Eq. (10) as follows,

$$h_4 = f_4 \odot (1 + E_4). \tag{10}$$

These fused feature vectors are computed respectively and semantically guide the domain-invariant feature learning at instance-level alignment. The adversarial loss of domain discriminator D_4 is defined as follows,

$$\begin{aligned}\mathcal{L}_4 = &\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_j \log(D_4(F_4(h_i^s))_j)^2 \\ &+ \frac{1}{N_t}\sum_{i=1}^{N_t}\sum_j \log(1 - D_4(F_4(h_i^t))_j)^2.\end{aligned} \tag{11}$$

In addition, we deploy two other auxiliary networks F_5 and F_6 at the middle and high levels of the backbone network. Its uncertainty-aware context features and adversarial losses, $h_5, \mathcal{L}_5$, $h_6, \mathcal{L}_6$, are respectively derived in a similar manner. These vectors h_4, h_5, h_6 are then concatenated with instance-level features f_{ins} which yields new instance-level features augmented, denoted as f_{aug}. The instance-level adversarial loss is defined as follows,

$$\begin{aligned}\mathcal{L}_{ins} = &-\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_j \log(D_{ins}(f_{aug}^s)_j) \\ &- \frac{1}{N_t}\sum_{i=1}^{N_t}\sum_j \log(1 - D_{ins}(f_{aug}^t)_j).\end{aligned} \tag{12}$$

3.4 Overall Objective Function

The overall losses of the proposed method compose of detection losses and adversarial losses. Detection losses denoted as $\mathcal{L}_{det}$ defined in Eq. (14) include object classification loss $\mathcal{L}_{cls}$ and bounding box regression loss $\mathcal{L}_{reg}$. The adversarial training losses denoted as $\mathcal{L}_{adv}$ are composed of as follows,

$$\mathcal{L}_{adv} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4 + \mathcal{L}_5 + \mathcal{L}_6 + \mathcal{L}_{ins}, \tag{13}$$

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{reg}. \tag{14}$$

The overall objective function is defined as follows,

$$\max_{D_i}\min_{F_i} \mathcal{L}_{det} + \lambda \cdot \mathcal{L}_{adv}, \tag{15}$$

where D_i represents multi-level domain classifiers D_1, D_2, D_3, D_4, D_5, D_6, D_{ins}; F_i represents multi-level feature extractors F_1, F_2, F_3, F_4, F_4, F_5, F_6; and λ is a hyperparameter controlling influence degree of adversarial training losses and detection losses.

4 EXPERIMENTS

In this section, we present the experiment setting in detail. Next, we evaluate our adaptation framework on various cross-domain scenarios including learning from synthetic data, domain gap in adverse weather, and real to artistic domain adaptation. In addition, ablation analysis is conducted to evaluate each proposed modules.

4.1 Datasets

Datasets. To set up different adaptation scenarios, we utilize six public datasets in our experiments including Cityscapes [4], Foggy Cityscapes [23], SIM10k [15], PASCAL VOC 0712 (2007-2012) [6], Clipart1k [14] and Watercolor [14].

Cityscapes [4] is collected from various scenarios of outdoor street scenes in different cities. These images are recorded in normal weather conditions by video cameras mounted on cars. It consists of 2,975 images in the training set and 500 images for the testing set. **Foggy Cityscapes** [23] is synthesized from the images in the Cityscapes dataset. We select the shortest range of visibility in our experiment. **SIM10k** [15] is a synthetic dataset that is rendered from the computer game Grand Theft Auto V (GTA V). It consists of 10,000 images of street scenes with 58,701 bounding box annotations for cars. **PASCAL VOC 0712** [6] is a well-known dataset containing bounding boxes for 20 object categories. Following the evaluation protocol previous studies, we combine PASCAL VOC 2007 and 2012 training and validation sets for training sets (16,551 images in total) for experiments. **Clipart1k** [14] is a collection of 1k comical images. This dataset includes the same 20 object categories as PASCAL VOC but presents a significant domain gap. **Watercolor** [14] is a collection of 2k artistic images. This dataset contains 6 instance categories in common with PASCAL VOC.

4.2 Implementation Details

Following the evaluation protocol in [3, 22], we adopt Faster-RCNN [21] and employ the VGG-16 [25] or ResNet-101 [11] as the backbone networks in which their weights are pre-trained on ImageNet. We follow the setting in [22] and use ResNet-101 as the backbone model for the dissimilar domain-shift adaptation scenarios from PASCAL VOC [6] to Clipart1k [14] and PASCAL VOC [6] to Watercolor [14]. For other domain-adaptive experiments, we use the pre-trained weights of VGG-16 to fine-tune the object detection model. Stochastic gradient descent (SGD) is selected as the optimizing method for the training process. The hyperparameter λ is set to 0.1 for SIM10k $\rightarrow$ Cityscape and $\lambda = 1$ for all other scenarios following the evaluation setting from prior work. To evaluate the adaptation performance, we report mean average precision (mAP) with IOU threshold of 0.5 for all the experiments.

4.3 Baselines

We compare our proposed method with the original Faster-RCNN [21] and a variety of recent state-of-the-art domain-adaptive detection methods as follows:

Figure 3: Demonstrations of the detection results on the target images, the domain-adaptive task from Cityscapes → Foggy Cityscapes. First row by SWDA (CVPR' 19). Second row by Ours. Best view in color.

Figure 4: Demonstrations of the uncertainty-aware attentional feature maps on the adaptation scenario from Cityscapes (Bottom) → Foggy Cityscapes (Top). Best view in color.

- **Domain Adaptive Faster-RCNN (DA-Faster) [3]** - A pioneer work that deploys domain classifiers via adversarial learning to tackle domain-shift problems at image and instance levels.
- **Strong Weak Distribution Alignment (SWDA) [22]** - Strong local image-level alignment and weak global alignment are applied.
- **Multi-Adversarial Faster-RCNN (MAF) [12]** - A hierarchical domain feature alignment and a weighted gradient reversal layer applied to penalize the large weights.
- **Selective Cross-Domain Alignment (SCDA) [33]** - Performing alignment at region level by clustering instance-level features to learn the discriminatory regions.
- **Domain Diversification and Multi-Domain-Invariant Representation Learning (DD-MRL) [17]** - Diversifying distribution of source domain and learn domain-invariant representation.
- **Hierarchical Transferability Calibration Network (HTCN) [2]** - Utilizing CycleGAN to generate interpolated images, context-aware vectors and local feature masks for calibrating the transferability and discriminability.
- **Image-level Categorical Regularization and Categorical Consistency Regularization (ICR-CCR) [30]** - Obtaining crucial image regions and hunting for hard-aligned instances.
- **Weak Self-Training and adversarial Background Score Regularization (WST-BSR) [16]** - Reducing the adverse effects of inaccurate pseudo-labels and facilitating the network to learn discriminative features.

Parts of the results of the above methods are cited from their papers accordingly.

4.4 Experimental Results

In this section, we present experimental results and compare our method's performance with baselines on various domain-adaptive scenarios.

Weather Adaptation. The results of the adaptation from Cityscapes to Foggy Cityscapes are reported in Table 1. The Source-Only method means no adaptation applied. Specifically, our method achieves a new state-of-the-art mean average precision (mAP) which significantly boosts the performance of cross-weather adaptation by 20.2% mAP over the Source-Only baseline (from 20.3% to 40.5%). Furthermore, our proposed framework outperforms all the recent state-of-the-art methods. In particular, it is worth noting that in the HTCN method [2], the CycleGAN is utilized to generate interpolation image samples for source and target domain, and then train their model with original and synthesized images. In terms of computational perspective, these additional steps are considerably more expensive than our method in which authentic images from

Table 1: Results on adaptation from Cityscapes to Foggy-Cityscapes. Average precision (%) is evaluated on target images.

Methods	Person	Rider	Car	Truck	Bus	Train	Motorbike	Bicycle	mAP
Source Only [21]	24.1	33.1	34.3	4.1	22.3	3.0	15.3	26.5	20.3
DA-Faster (CVPR' 18) [3]	25.5	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
SCDA (CVPR' 19) [33]	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
MAF (ICCV' 19) [12]	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
SWDA (CVPR' 19) [22]	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
DD-MRL (CVPR' 19) [17]	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
ICR-CCR (CVPR' 20) [30]	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
HTCN (CVPR' 20) [2]	33.2	47.5	47.9	**31.6**	**47.4**	40.9	32.3	**37.1**	39.8
MEAA (Ours)	**34.2**	**48.9**	**52.4**	30.3	42.7	**46.0**	**33.2**	36.2	**40.5**
Oracle [22]	33.2	45.9	49.7	35.6	50.0	37.4	34.7	36.2	40.3

Table 2: Results on adaptation from PASCAL VOC to Clipart Dataset. Average precision (%) is evaluated on target images.

Methods	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	mbike	prsn	plnt	sheep	sofa	train	tv	mAP
Source Only [21]	**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 (CVPR' 18) [3]	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
WST-BSR (ICCV' 19) [16]	28.0	**64.5**	23.9	19.0	21.9	64.3	**43.5**	16.4	**42.2**	25.9	**30.5**	7.9	25.5	67.6	54.5	36.4	10.3	**31.2**	57.4	32.5	35.7
SWDA (CVPR' 19) [22]	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
ICR-CCR (CVPR' 20) [30]	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 (CVPR' 20) [2]	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
MEAA (Ours)	31.3	53.5	**38.0**	17.8	38.5	**69.9**	38.2	**23.8**	38.3	58.1	14.6	18.1	33.8	**88.1**	60.3	42.1	7.8	30.8	**61.1**	**58.7**	**41.1**

Table 3: Results on adaptation from PASCAL VOC to WaterColor Dataset. Average precision (%) is evaluated on target images.

Methods	bike	bird	car	cat	dog	person	mAP
Source Only [21]	68.8	46.8	37.2	32.7	21.3	60.7	44.6
DA-Faster (CVPR' 18) [3]	75.2	40.6	48.0	31.5	20.6	60.0	46.0
WST-BSR (ICCV' 19) [16]	75.6	45.8	49.3	34.1	30.3	64.1	49.9
SWDA (CVPR' 19) [22]	**82.3**	**55.9**	46.5	32.7	35.5	**66.7**	53.3
MEAA (Ours)	81.0	53.2	**54.0**	**40.1**	**39.2**	65.3	**55.5**

Table 4: Results on Sim10k to Cityscapes (%). The backbone network is VGG-16.

Methods	AP on *car*
Source Only [21]	34.6
DA-Faster (CVPR' 18) [3]	38.9
SWDA (CVPR' 19) [22]	40.1
MAF (ICCV' 19) [12]	41.1
HTCN (CVPR' 20) [2]	**42.5**
MEAA (Ours)	42.0

the source and target domain are directly fed into our model.

Dissimilar Domain Adaptation. We carry out experiments from real source images to artistic target images which are considered as the dissimilar domain adaptation. Table 2 shows the results of the adaptation from **PASCAL VOC** to **Clipart1k**, which manifests that our proposed method considerably improves the performance over the Source-Only [21], DA-Faster [3] and SWDA [22] in terms of mAP by 13.3%, 21.3% , and 3%, respectively. Furthermore, our approach also outperforms the recent state-of-the-art ICR-CCR [30] and HTCN [2]. Results on the adaptation task from **PASCAL VOC** to **Watercolor** are presented in Table 3, which demonstrate that our method improves the state-of-the-art performance. The enhancement of the challenging dissimilar domain adaptation clearly validates the robustness of our method.

Adaptation from synthetic to real images. The results of the adaptation from SIM10k to Cityscapes are shown in Table 4. Our method's performance boosts the average precision by 7.4% over the Source-Only [21] baseline. Moreover, it outperforms most of the state-of-the-art comparison methods and provides comparable performance compared to the HTCN [2]. It is worthy to note that the HTCN model's inputs are not only original target and source domain images but also CycleGAN-synthesized target and source images which are considerably more expensive than our approach with regard to computational cost.

Table 5: Ablation study on adaptation from Cityscapes to Foggy-Cityscapes. Average precision (%) is evaluated on target images.

Methods	Person	Rider	Car	Truck	Bus	Train	Motorbike	Bicycle	mAP
Source Only	24.1	33.1	34.3	4.1	22.3	3.0	15.3	26.5	20.3
MEAA without LUAA	**41.5**	34.0	52.0	**33.0**	34.0	46.9	39.3	26.5	38.4
MEAA without MUCA	41.3	35.4	52.1	31.5	34.0	**47.0**	**44.1**	24.7	38.8
MEAA (full)	34.2	**48.9**	**52.4**	30.3	**42.7**	46.0	33.2	**36.2**	**40.5**
Oracle (Target only) [22]	33.2	45.9	49.7	35.6	50.0	37.4	34.7	36.2	40.3

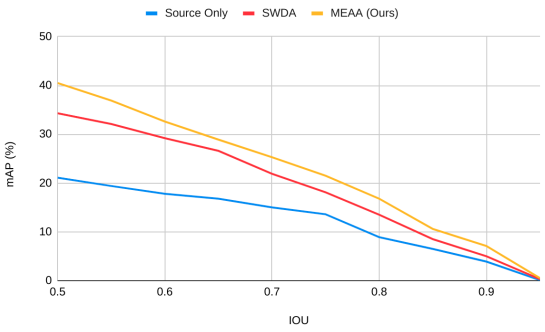


Figure 5: Demonstrations of the performance with respect to different IOU setting on the adaptation scenario from Cityscapes (Bottom) → Foggy Cityscapes (Top). Best view in color.

4.5 Visualization and Analysis

In this section, we present the ablation experiments and visualization examples to investigate the effect of different modules on our model.

Ablation Study. Table 5 shows that if one of the modules is removed, the performance will correspondingly degrade. The performance in Table 5 implies that the two new modules are effective and complementary for each other. Furthermore, Table 5 shows that the evaluation metrics of MEAA (full) for different objects are all very close to the performance of the Oracle (the last row of Table 5). For example, in categories Person and Motorbike, the average precisions of MEAA are respectively 34.2% and 33.2%, which are both reasonable results compared to that of the Oracle (33.2% and 34.7%) and also higher than that of other methods with empirically acceptable margins (in Table 1). Thus, although the performance of Person and Motorbike degrade when combining LUAA and MUCA for a few categories, the MEAA (full) results in a better mAP in general. We believe that the instability of the model combining LUAA and MUCA comes from the performance tradeoff between some pairs of class categories that have a similar structure such as Person and Rider, Motorbike and Bicycle. It is promising to investigate a better way of striking a better balance.

Influence of IOU thresholds. To evaluate the accuracy and robustness of the object classification and bounding box regression, in figure 5, we illustrate the performance of our model compared with Source-Only and SWDA [22] with varying values of IOU thresholds. The results show that the proposed method considerably outperforms the comparison methods which indicates that our model is accurate and robust of object classification and bounding box localization.

Visualization of Uncertainty-Aware Feature Maps. In figure 4, we visualize the heatmap of local uncertainty-aware attentional feature maps from the local uncertainty-aware attentional alignment (LUAA) module. The figure illustrates that the MEAA pay more attention to heterogeneous instances of interest (*e.g.* car, person) with higher entropy rather than on homogeneous regions (*e.g.* road, wall). This verifies our hypothesis that objects of interest are more likely with high entropy due to its varying structure and texture. As a result, the uncertainty-aware attentional feature maps facilitate the model semantically focusing on objects of interest to better realize domain-invariant representation in feature space.

Detection Result Examples. We illustrate several qualitative results of our MEAA model. Figure 3 visualizes the domain-adaptive detection results on scenarios of normal-to-foggy and real images to artistic images, Cityscapes → Foggy Cityscapes. The figure clearly demonstrates that the MEAA is capable of detecting challenging objects with accurate bounding box localization which leads the proposed model to yield more robust detection than SWDA [22]. This implies our proposed model substantially close domain-gap distribution between source images and the target images.

5 CONCLUSION AND FUTURE WORK

In this work, we present an end-to-end approach for improving the domain-adaptive detection performance of object detectors in the unsupervised adaptation protocol. Specifically, we exploited uncertainty measurement of input images by utilizing domain classifiers at multiple-level of the network and the uncertainty at image and instance levels which facilitate the model to pay more attention to hard-to-align instances and images. The extensive experiments demonstrate that our proposed method improves state-of-the-art performance. In the future, we plan to explore domain adaptation beyond 2D object detection by investigating how to apply the uncertainty-aware scheme to improve domain-adaptive 3D object detectors.

ACKNOWLEDGEMENTS

We are grateful to the National Center for High-performance Computing for computer time and facilities. This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST-109-2221-E-009-114-MY3, MOST-109-2634-F-009-018, MOST-108-2622-E-009-026-CC2 and MOST-109-2218-E-009-025.

REFERENCES

[1] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. 2019. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1900–1909.

[2] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. 2020. Harmonizing Transferability and Discriminability for Adapting Object Detectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3339–3348.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.

[7] Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495* (2014).

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[9] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Zhenwei He and Lei Zhang. 2019. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 6668–6677.

[13] J Hoffman, D Wang, F Yu, and T Darrell. [n.d.]. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv 2016. *arXiv preprint arXiv:1612.02649* ([n. d.]).

[14] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5001–5009.

[15] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. 2016. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983* (2016).

[16] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. 2019. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 6092–6101.

[17] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. 2019. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 12456–12465.

[18] Shuang Li, Chi Harold Liu, Binhui Xie, Limin Su, Zhengming Ding, and Gao Huang. 2019. Joint Adversarial Domain Adaptation. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery, New York, NY, USA, 729–737. https://doi.org/10.1145/3343031.3351070

[19] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. 2015. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791* (2015).

[20] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[22] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. 2019. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6956–6965.

[23] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision* 126, 9 (2018), 973–992.

[24] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. 2018. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3752–3761.

[25] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[26] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. 2018. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7472–7481.

[27] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7167–7176.

[28] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2517–2526.

[29] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S. Yu. 2018. Visual Domain Adaptation with Manifold Embedded Distribution Alignment. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. Association for Computing Machinery, New York, NY, USA, 402–410. https://doi.org/10.1145/3240508.3240512

[30] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. 2020. Exploring Categorical Regularization for Domain Adaptive Object Detection. *arXiv preprint arXiv:2003.09152* (2020).

[31] Yuan Yao, Yu Zhang, Xutao Li, and Yunming Ye. 2019. Heterogeneous Domain Adaptation via Soft Transfer Network. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1578–1586.

[32] Xiheng Zhang, Yongkang Wong, Mohan S. Kankanhalli, and Weidong Geng. 2019. Unsupervised Domain Adaptation for 3D Human Pose Estimation. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery, New York, NY, USA, 926–934. https://doi.org/10.1145/3343031.3351052

[33] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. 2019. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 687–696.

[34] Junbao Zhuo, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2017. Deep Unsupervised Convolutional Domain Adaptation. In *Proceedings of the 25th ACM International Conference on Multimedia (MM '17)*. Association for Computing Machinery, New York, NY, USA, 261–269. https://doi.org/10.1145/3123266.3123292