Exploring Categorical Regularization for Domain Adaptive Object Detection (Supplementary Materials)

In the supplementary materials, we present more details of computing domain distance, and more experimental studies and visualization understanding about our categorical regularization framework.

- We briefly review the formal definition of Earth Mover's Distance (EMD) [5] used in our quantitative analyses (cf. Section 4.3 of the paper), and describe our calculation process of EMD for measuring the domain distance of the source and target detection datasets. We also provide explanations about why we adopt the instance-based (rather than image-based) domain representation for detection datasets.
- We conduct a series of experiments for the strong-weak aligned Faster R-CNN (SW-Faster) [6] to investigate the effects of additional plain instance-level alignment models. We do this because the original pipeline of SW-Faster has no integrated instance-level alignment model, while our categorical consistency regularization (CCR) module relies on the existence of instance-level alignment.
- We give more visualization examples about the baseline SW-Faster method [6] and our SW-Faster-ICR-CCR method to demonstrate the effectiveness of our categorical regularization framework. We include more detection examples on three target datasets (*i.e.*, Foggy Cityscapes [7], BDD100k [9], Clipart1k [3]), and show more heatmaps on the sourced Cityscapes [2] dataset and the targeted Foggy Cityscapes dataset.

1. Earth Mover's Distance as Domain Distance/Similarity for Detection Datasets

In our experiments (cf. Section 4.3 of the paper), we employ Earth Mover's Distance (EMD) [5] as a quantitative measure for the domain distance between source and target detection datasets. A high EMD indicates a low similarity between domains, and conversely, a low EMD indicates a high similarity. EMD is well suited for our case as it can be directly calculated by the samples from both domains, while other measures such as Kullback-Leibler (KL) divergence require additional modeling process for domain distributions.

In Section 4.3 of the paper, we demonstrate that our method can achieve higher domain similarity (lower EMD) than the baseline methods at the *instance* level. This result quantitatively validates that our method can better align objects of interest of both domains, and thus can boost the performance of domain adaptive detection. In this following, we first briefly review the formal definition of EMD, and then explain why we utilize the instance samples rather than the whole image to represent the domain.

1.1. Formal Definition of EMD and Its Application to Classification Datasets

Earth Mover's Distance (EMD) [4, 5] defines a distance metric for two sets (*e.g.*, the source and target images), using the sample clusters (*e.g.*, images of the same category) in these two sets. In the context of domain adaptation, we refer to these two sets as source set and target set. Formally, let $\mathcal{S} = \{(\mathbf{s}_i, w_{\mathbf{s}_i})\}_{i=1}^{m}$ be the source set with m clusters, where $\mathbf{s}_i$ is the cluster representation, and $w_{\mathbf{s}_i}$ is the weight of the cluster which corresponds to the normalized cluster size; let $\mathcal{T} = \{(\mathbf{t}_j, w_{\mathbf{t}_j})\}_{j=1}^{n}$ be the target set with j clusters; and $\mathbf{D} = [d_{ij}]$ is the ground distance matrix where d_{ij} is the ground distance between clusters $\mathbf{s}_i$ and $\mathbf{t}_j$. The EMD between the source set $\mathcal{S}$ and target set $\mathcal{T}$ measures the least amount of "work" needed to transform one set into the other, and can be formulated as a weighted distance of all cluster distances as follows

$$EMD(\mathcal{S}, \mathcal{T}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}, \tag{1}$$

where f_{ij} is the flow between source cluster s_i and target cluster t_j. The optimal flow f_{ij} corresponds to the least amount of total work by solving the EMD optimization problem. We refer the interested readers to [5] for the explanation of flow f_{ij} and the objective function for optimizing f_{ij}.

Given above definition of EMD, we can conveniently calculate the domain distance between two image classification datasets in three steps. The first step is to divide the images of each domain into clusters based on image categories, then extract the representation for each category by averaging the GAP (global average pooling) features of all images of this category, and calculate the normalized weight for each category based on the number of images of this category. The second step is to compute the distance matrix $\mathbf{D} = [d_{ij}]$ for the category pairs from different domains, where d_{ij} is the ℓ_2 distance between $\mathbf{s}_i$ and $\mathbf{t}_j$, *i.e.*, the representation of category i in source domain and the representation of category j in target domain. The last step is to find the optimal flow f_{ij} by solving the EMD optimization problem, and calculate the EMD between source and target domains according to Equation 1.

1.2. EMD with Instance-based Domain Representation for Detection Datasets

As aforementioned, it is straightforward to apply EMD to measure the domain distance of source and target datasets for image classification. However, while for object detection datasets, it is unrealistic to divide the images into clusters based on image categories, since each image typically contains multiple objects of different categories.

To make EMD tractable for detection datasets, we employ an instance-based domain representation. We extract the ground truth instances from each image, where each instance is associated with a single object label. This allows us to convert a detection dataset into an image classification dataset, by treating each instance as an individual image. With the instance-based domain representation, we can apply EMD to measure the domain distance of the source and target detection datasets. Since the domain adaptive detection task relies heavily on the alignment of objects of interests, instance-level (object-level) domain distance can better reflect the domain adaptation performance associated with detection. As demonstrated in Section 4.3, using the instance-level EMD as a metric for domain distance, our SW-Faster-ICR and SW-Faster-CCR detectors perform better than the baseline SW-Faster detector [6].

2. Investigation about Additional Instance-level Alignment for SW-Faster [6]

While our categorical consistency regularization (CCR) module is built upon the instance alignment model, the original strong-weak aligned Faster R-CNN (SW-Faster) [6] has no integrated instance-level alignment model. In order to further improve SW-Faster with proposed CCR, we add an instance-level alignment model to SW-Faster, which is same to that of the DA Faster R-CNN [1]. To fairly validate the effectiveness of our CCR module, we conduct a series of experiments for SW-Faster, by integrating the plain instance-level alignment module and our CCR enhanced instance-level alignment module, respectively. Using the same experimental settings with Section 4 of the paper, these experiments cover three domain adaptation scenarios, *i.e.*, weather adaptation, scene adaptation and dissimilar domain adaptation.

Table 1, Table 2 and Table 3 of the supplementary materials show the comparison results, where "SW-Faster-Instance" and "**SW-Faster-ICR**-Instance" denote SW-Faster and SW-Faster-ICR integrated with additional plain instance alignment model, respectively. We observe that integrating SW-Faster with plain instance alignment does not obtain obvious accuracy gains for weather adaptation and scene adaptation, but brings considerable benefits for dissimilar domain adaptation. We speculate that it is because the strong local alignment model in SW-Faster plays a similar role of instance alignment, and thus reduces the effect of additional instance alignment module for similar domain adaptation. While for dissimilar domain adaptation, the strong local alignment in SW-Faster faces the risk of overfitting the non-transferable source backgrounds, but the instance-level alignment has the advantage of aligning possible objects of interests.

For SW-Faster-ICR, our CCR enhanced instance alignment consistently outperforms the plain instance alignment in all experiments, although the latter also achieves slight improvement over the baseline methods. On the other hand, due to the existence of strong local alignment, the improvements over the plain instance alignment by our CCR is not as significant as that for DA-Faster which has an integrated instance alignment model (see Section 4 of the paper).

3. More Visualization Examples

To demonstrate the effectiveness of our categorical regularization framework more intuitively, we give more visualization examples of the state-of-the-art SW-Faster [6] detector and our SW-Faster-ICR-CCR detector, including more detection examples on three target datasets (*i.e.*, Foggy Cityscapes [7], BDD100k [9], Clipart1k [3]), and more heatmaps on the sourced Cityscapes [2] dataset and the targeted Foggy Cityscapes dataset.

Table 1. Weather Adaptation: Comparisons of SW-Faster [6] augmented with different **instance-level alignment** methods.

Method	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
SW-Faster [6]	32.3	42.2	47.3	23.7	41.3	27.8	28.3	35.4	34.8
SW-Faster-Instance	32.5	42.0	45.6	20.6	41.6	32.7	27.6	32.7	34.4
SW-Faster-ICR (Ours)	**33.1**	**44.2**	48.8	**27.7**	44.9	27.9	29.4	**36.2**	36.5
SW-Faster-ICR-Instance	32.6	43.7	49.1	26.4	**45.2**	32.6	29.6	35.1	36.8
SW-Faster-ICR-CCR (Ours)	32.9	43.8	**49.2**	27.2	45.1	**36.4**	**30.3**	34.6	**37.4**

Table 2. Scene Adaptation: Comparisons of SW-Faster [6] augmented with different **instance-level alignment** methods.

Methods	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
SW-Faster [6]	30.2	29.5	45.7	15.2	18.4	-	17.1	21.2	25.3
SW-Faster-Instance	31.1	29.5	**46.4**	16.9	17.3	-	16.0	23.3	25.8
SW-Faster-ICR (Ours)	30.9	31.2	45.6	15.9	18.4	-	**19.3**	23.7	26.4
SW-Faster-ICR-Instance	31.3	30.0	46.3	19.2	**19.6**	-	16.5	**23.8**	26.7
SW-Faster-ICR-CCR (Ours)	**31.4**	**31.3**	46.3	**19.5**	18.9	-	17.3	**23.8**	**26.9**

Table 3. Dissimilar Domain Adaptation: Comparisons of SW-Faster [6] augmented with different **instance-level alignment** methods.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
SW-Faster [6]	**29.2**	53.1	30.2	24.4	**41.4**	52.5	34.6	**14.0**	36.3	43.5	17.6	**16.6**	**33.4**	**78.1**	59.1	42.1	15.8	24.9	45.5	43.7	36.8
SW-Faster-Instance	27.9	52.0	31.5	23.2	37.2	58.3	36.3	13.6	**39.2**	**54.6**	17.8	15.0	32.5	71.8	**63.9**	**47.3**	16.9	24.8	46.0	38.5	37.4
SW-Faster-ICR (Ours)	25.2	54.0	31.7	23.4	40.3	65.8	35.4	12.1	37.6	48.1	**18.6**	14.2	31.3	73.6	59.9	46.5	19.5	**25.9**	46.0	**45.6**	37.7
SW-Faster-ICR-Instance	27.6	52.6	29.9	23.5	38.2	**68.4**	**37.9**	11.3	37.8	48.0	16.6	12.9	33.1	75.3	60.5	46.4	20.1	24.5	**50.3**	42.6	37.9
SW-Faster-ICR-CCR (Ours)	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**

3.1. Detection Examples

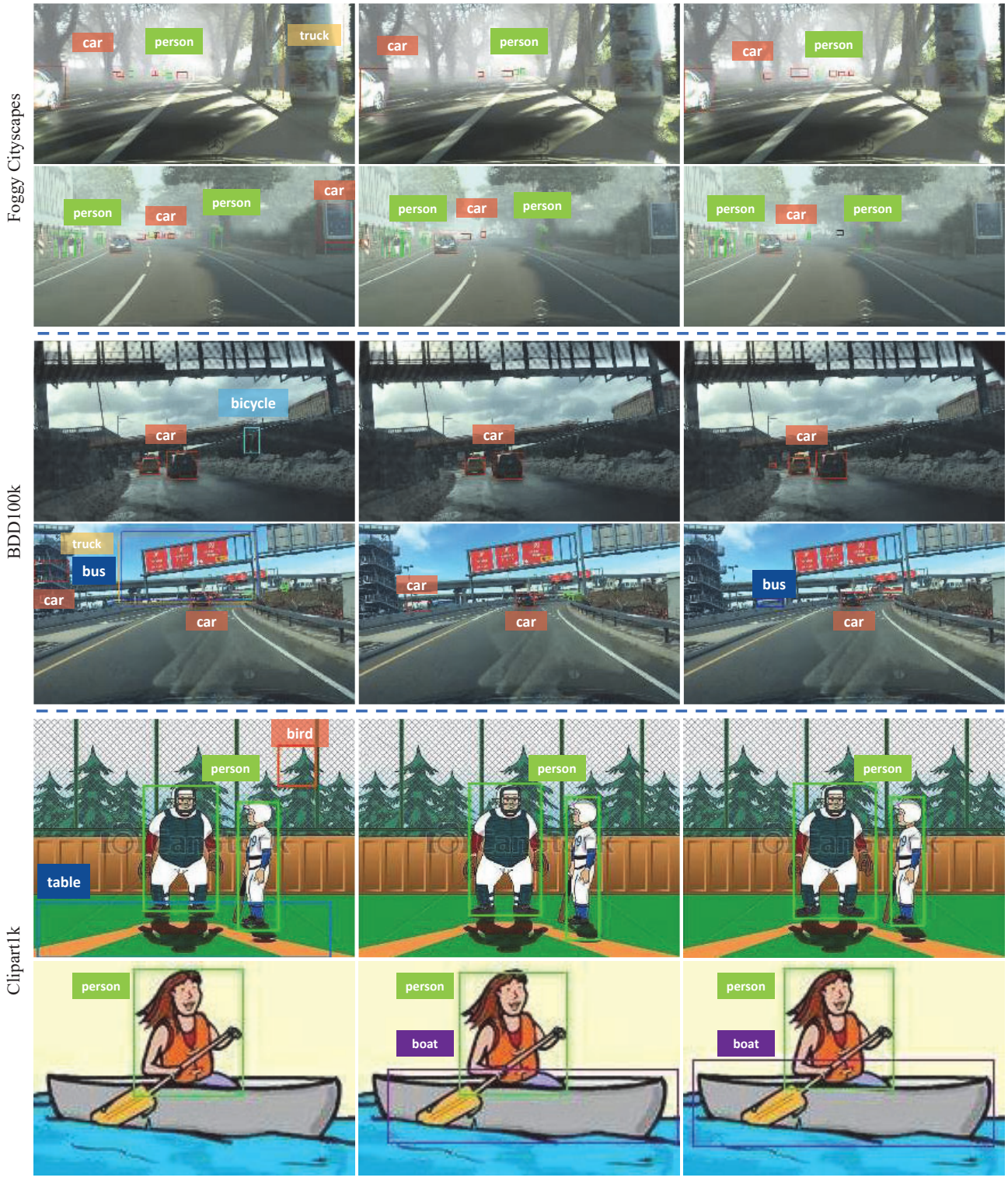
Figure 1 of the supplementary materials shows some detection examples on three target datasets, *i.e.*, Foggy Cityscapes [7], BDD100k [9] and Clipart1k [3], with comparisons between our SW-Faster-ICR-CCR method and the baseline SW-Faster [6] method. We also provide the ground truth annotations as the references.

These example images demonstrate that our categorical regularization framework can significantly improve SW-Faster [6], and produce more accurate detections under complex environments and large domain shifts. In particular, as shown by the first four rows in Figure 1 of the supplementary materials , our method can reduce the false positive detections of SW-Faster to a large extent. We speculate that this is because the original domain adaptation components in SW-Faster face the risk of overfitting the source backgrounds due to the RoI-based training of Faster R-CNN, while our image-level categorical regularization can reduce this overfitting risk to a certain extent. Since there is no background supervision signals involved in the training of our image-level multi-label classifier, the multi-label classification loss will drive the learning process to focus on discriminative features of foreground objects of interest.

3.2. Heatmap Examples

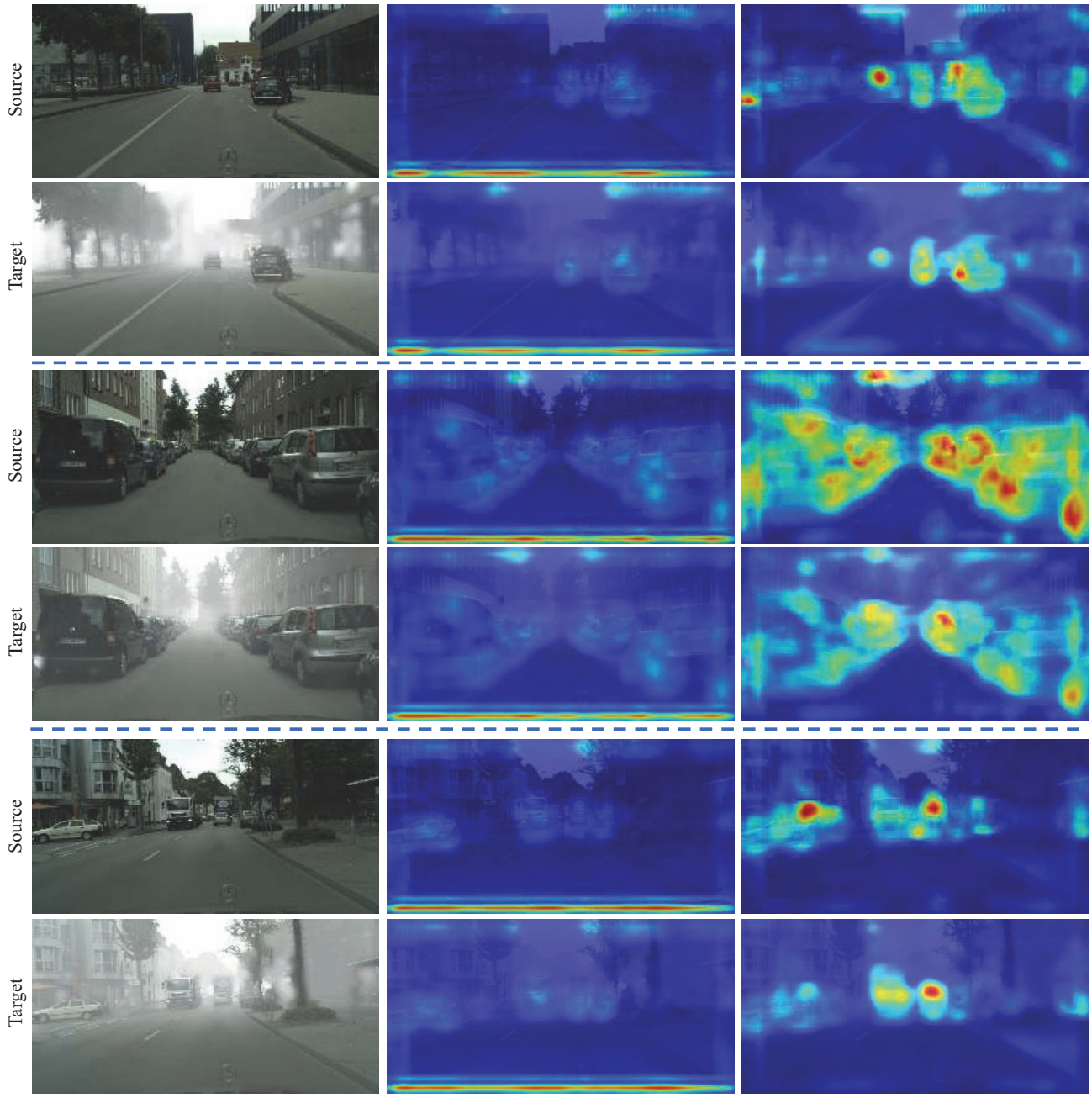
In Figure 2 of the supplementary materials , we give more examples of the heatmaps on source images from Cityscapes [2] and target images from Foggy Cityscapes [7]. These heatmap examples are produced by the detection backbone networks (*i.e.*, the VGG-16 [8] model) of the baseline SW-Faster [6] detector and our SW-Faster-ICR-CCR detector. Since our regularization framework enables more accurate alignment for crucial regions and important instances, it can assist the backbone network to activate the main objects of interest more accurately in *both domains*.

In particular, we observe an interesting phenomenon that SW-Faster trends to produce strong activations on the bottom of an image. We speculate that this is because the bottom of a target image is less affected by the foggy weather, and thus can be better aligned by the weak global alignment model of SW-Faster. While the objects of interest on both domains can be hardly aligned by the weak global model of original SW-Faster, our regularization framework enables much better alignment due to it ability of finding the crucial regions and important instances.



(a) SW-Faster (b) SW-Faster-ICR-CCR(**Ours**) (c) Ground Truth

Figure 1. Detection examples from three target datasets, from top to bottom: Foggy Cityscapes [7], BDD100k [9], and Clipart1k [3]. Our categorical regularization framework enables SW-Faster [6] to produce more accurate detection results with large domain shifts.



(a) Source/Target Images (b) SW-Faster (c) SW-Faster-ICR-CCR(**Ours**)

Figure 2. (a): Example images from Cityscapes [2] (source) and Foggy Cityscapes [7] (target). **(b)**: Heatmaps by the backbone network (VGG-16 [8]) of SW-Faster [6]. **(c)**: Heatmaps by the backbone network of SW-Faster trained *with* our categorical regularization framework. Our regularization framework enables more accurate alignment for crucial regions and important instances, and thus can assist the backbone network to activate the main objects of interest more accurately in *both domains*, and lead to better adaptive detection performance.

References

[1] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive Faster R-CNN for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 2

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 1, 2, 3, 5

[3] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018. 1, 2, 3, 4

[4] Svetlozar T Rachev. The Monge–Kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, 29(4):647–676, 1985. 1

[5] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000. 1, 2

[6] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 1, 2, 3, 4, 5

[7] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018. 1, 2, 3, 4, 5

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, pages 1–8, 2015. 3, 5

[9] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018. 1, 2, 3, 4