# Supplementary Material for PM-DETR

Anonymous Author(s)
Submission Id: 1468

## A ADDITIONAL THEORETICAL ANALYSIS

Unfortunately, due to limitations in the available text space, we are unable to provide a comprehensive list of graphs and formulas to fully demonstrate the validity of our method. Therefore, we will now offer a more detailed explanation to address this constraint.

The motivation behind most previous papers [5, 8, 11, 14, 15] is derived from the theory of domain adaptation [1, 2]. This theory suggests that an effective representation for cross-domain transfer should be one that prevents algorithms from discerning the domain of origin of input observations. For domain adapative task, suppose there are source labeled data $S$ sampled from distribution $\mathcal{D}_S^{\mathrm{p}}$ and target unlabeled data $T$ sampled from distribution $\mathcal{D}_T^{\mathrm{q-p}}$, Our objective is to train a model $D \in \mathcal{H}$ (where $\mathcal{H}$ denotes the hypothetical space) for minimizing the error $\varepsilon_T$. To accomplish this, we utilize supervised learning on the labeled data $S$ and unsupervised learning on the unlabeled data $T$. The model $D$ should be capable of effectively leveraging the labeled data while leveraging the inherent structure and patterns present in the unlabeled data to improve its performance on the target domain.

$$\min \varepsilon_T(D) = \min \Pr_{(x,y)\in D_T^{\mathrm{q-p}}}(D(x) \neq y) \tag{1}$$

A commonly held assumption among researchers is that the source risk can serve as a reliable estimate for the target risk when the underlying distributions are similar. Hence, there is a pressing need for an accurate approach to measure the distance between distributions. This principle is exemplified in the research conducted by Ben-David et al. [1, 2], the distance between $D_S$ and $D_T$ can be well characterized by $\mathcal{H}-divergence$:

$$d_{\mathcal{H}}(D_S, D_T) = 2 \sup_{D\sim\mathcal{H}} |\Pr_{x\sim\mathcal{D}_S}[D(x)=1] - \Pr_{x\sim\mathcal{D}_T}[D(x)=1]| \tag{2}$$

Indeed, $\mathcal{H}-divergence$ depends on the model space $\mathcal{H}$ and the underlying domain data distribution. When $S$ and $T$ are generated by sampling from their respective distribution, the empirical $\mathcal{H}-divergence$ provides an estimate of the dissimilarity between the distributions based on the observed data:

$$\hat{d}_{\mathcal{H}}(S,T) = 2(1 - \min_{D\in\mathcal{H}}[\frac{1}{p}\sum_{i=1}^{p}\mathcal{I}[D(x_i)=0] + \frac{1}{q-p}\sum_{j=p+1}^{q}\mathcal{I}[D(x_j)=1]]) \tag{3}$$

While $\mathcal{I}(\cdot)$ equals to one when $\cdot$ is true, and zero otherwise. Overall, Ben-David et al. [1, 2] show that $d_{\mathcal{H}}(D_S, D_T)$ is uppper bounded by its empirical estimate $\hat{d}_{\mathcal{H}}(D_S, D_T)$ plus a constant complexity term that depends on the VC dimension of $\mathcal{H}$. Let $\mathcal{H}$ be a hypothesis class of VC dimension $d$. With probablility $1-\sigma$ over the choice of samples $S \sim \mathcal{D}_S^{\mathrm{p}}$ and $T \sim \mathcal{D}_T^{\mathrm{p}}$, for every $D \in \mathcal{H}$:

$$\begin{aligned} \underset{D\in\mathcal{H}}{\varepsilon_T}(D) &\le \varepsilon_S(D) + \sqrt{\frac{4}{\mathrm{p}}(dlog\frac{2e\mathrm{p}}{d} + log\frac{4}{\sigma})} + \hat{d}_{\mathcal{H}}(S,T) \\ &+ 4\sqrt{\frac{1}{\mathrm{p}}(dlog\frac{2\mathrm{p}}{d} + log\frac{4}{\sigma})} + \beta = \sup \varepsilon_T(D) \end{aligned} \tag{4}$$

While $\beta \ge \inf_{D^*\in\mathcal{H}}[\varepsilon_T(D^*) + \varepsilon_S(D^*)]$. In the context of discriminating between source and target examples, the empirical $\mathcal{H}-divergence$ can be estimated using a proxy called the $\mathcal{A}-distance$, denoted as$\hat{d}_{\mathcal{A}} = 2(1-\epsilon)$, where $\epsilon$ represents the generalization error. Based on Eq. 4, during the training process, different approaches strive to promote the emergence of features that exhibit two important properties:(i) they are highly discriminative for the primary learning task on the source domain, and (ii) they are agnostic or invariant to the distribution shift between the domains.

However, previous methods often utilize use the same hypothetical space $\mathcal{H}$ for both the source and target domains, which inevitably introduces compromise error in each term of sup $\varepsilon_T(D)$. In the case of $\varepsilon_S(D)$, the parameters are influenced by unsupervised learning in the target domain, resulting in inferior performance. In the case of $\hat{d}_{\mathcal{H}}(S,T)$, imposing consistency constraints on the source and target domains is often ineffective due to small inter-class distances and large intra-class distances. For instance, SIGMA++ [7] demonstrate that employing agnostic structural dependencies fails to capture class variances, leading to sub-optimal outcomes. Moreover, $\beta$, as proven in [12], suggests that different domains maintain their unique characteristics. It is challenging to reach peak performance for either $\varepsilon_T(D^*)$ or $\varepsilon_S(D^*)$ with one single model $D$. Consequently, compromise error end-to-end impact sup $\varepsilon_T(D)$, the upper boundary of $\varepsilon_T(D)$ is not sufficiently tight. To overcome these limitations, our proposed method, PM-DETR, introduces the concepts of prompt domain memory (PDM) and prompt memory alignment (PMA) for decoupling $D \in \mathcal{H}$ to $D_S \in \mathcal{H}_S$ and $D_T \in \mathcal{H}_T$ respectively. Then Eq. 4 can be reformulated:

$$\begin{aligned} \underset{D_T\in\mathcal{H}_T}{\varepsilon_T}(D_T) &\le \varepsilon_S(D_S) + \sqrt{\frac{4}{\mathrm{p}}(dlog\frac{2e\mathrm{p}}{d} + log\frac{4}{\sigma})} + \hat{d}_{\mathcal{H}_S,\mathcal{H}_T}(S,T) \\ &+ 4\sqrt{\frac{1}{\mathrm{p}}(dlog\frac{2\mathrm{p}}{d} + log\frac{4}{\sigma})} + \beta(\mathcal{D}_S, \mathcal{D}_T) < \sup_{D\in\mathcal{H}} \varepsilon_T(D) \end{aligned} \tag{5}$$

Our approach elegantly resolves the common compromise error in domain adaptation problems through the utilization of prompt memory, which requires only a small number of parameters. Concretely, a hierarchical prompt domain memory (PDM) which constructs a long-term memory space aims to explore diversity domain-specific knowledge and fully learn the complex data distribution. This will greatly reduce $\beta$ and $\varepsilon_S(D_S)$ in $\varepsilon_{T\,D_T\in\mathcal{H}_T}(D_T)$. A prompt memory alignment (PMA) method aims to reduce the distribution

**Table 1: Performance comparison of methods published in 2023 conferences for weather adaptation, that is, from Cityscapes to Foggy Cityscapes. FRCNN and DefDETR are abbreviations for Faster R-CNN and Deformable DETR, respectively.**

| Method | Detector | Publication | person | rider | car | truck | bus | train | mcycle | bicycle | mAP | Gain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Two Stage :* | | | | | | | | | | | | |
| DA-AD [6] | FRCNN | WACV2023 | 51.2 | 39.1 | 54.3 | 31.6 | 36.5 | 46.7 | 48.7 | 30.3 | 42.3 | +13.8 |
| CMT [3] | FRCNN | CVPR2023 | 42.3 | 51.7 | 64.0 | 26.0 | 42.7 | 37.1 | 42.5 | 44.0 | 43.8 | +15.3 |
| *One Stage :* | | | | | | | | | | | | |
| ConfMix [8] | YOLOv5 | WACV2023 | 45.0 | 43.4 | 62.6 | 27.3 | 45.8 | 40.0 | 28.6 | 33.5 | 40.8 | +12.3 |
| AIRA-DA [10] | FCOS | LRA2023 | 43.6 | 46.7 | 62.1 | 27.8 | 44.0 | 37.0 | 29.9 | 38.4 | 41.2 | +12.7 |
| *Transformer based :* | | | | | | | | | | | | |
| Def DETR [16] (Source) | DefDETR | ICLR2021 | 37.7 | 39.1 | 44.2 | 17.2 | 26.8 | 5.8 | 21.6 | 35.5 | 28.5 | +00.0 |
| DA-DETR [15] | DefDETR | CVPR2023 | 49.9 | 50.0 | 63.1 | 24.0 | 45.8 | 37.5 | 31.6 | 46.3 | 43.5 | +15.0 |
| PM-DETR(Ours) | DefDETR | - | **47.8** | **50.2** | **64.7** | **26.5** | **47.2** | **39.6** | **32.4** | **46.1** | **44.3** | **+15.8** |

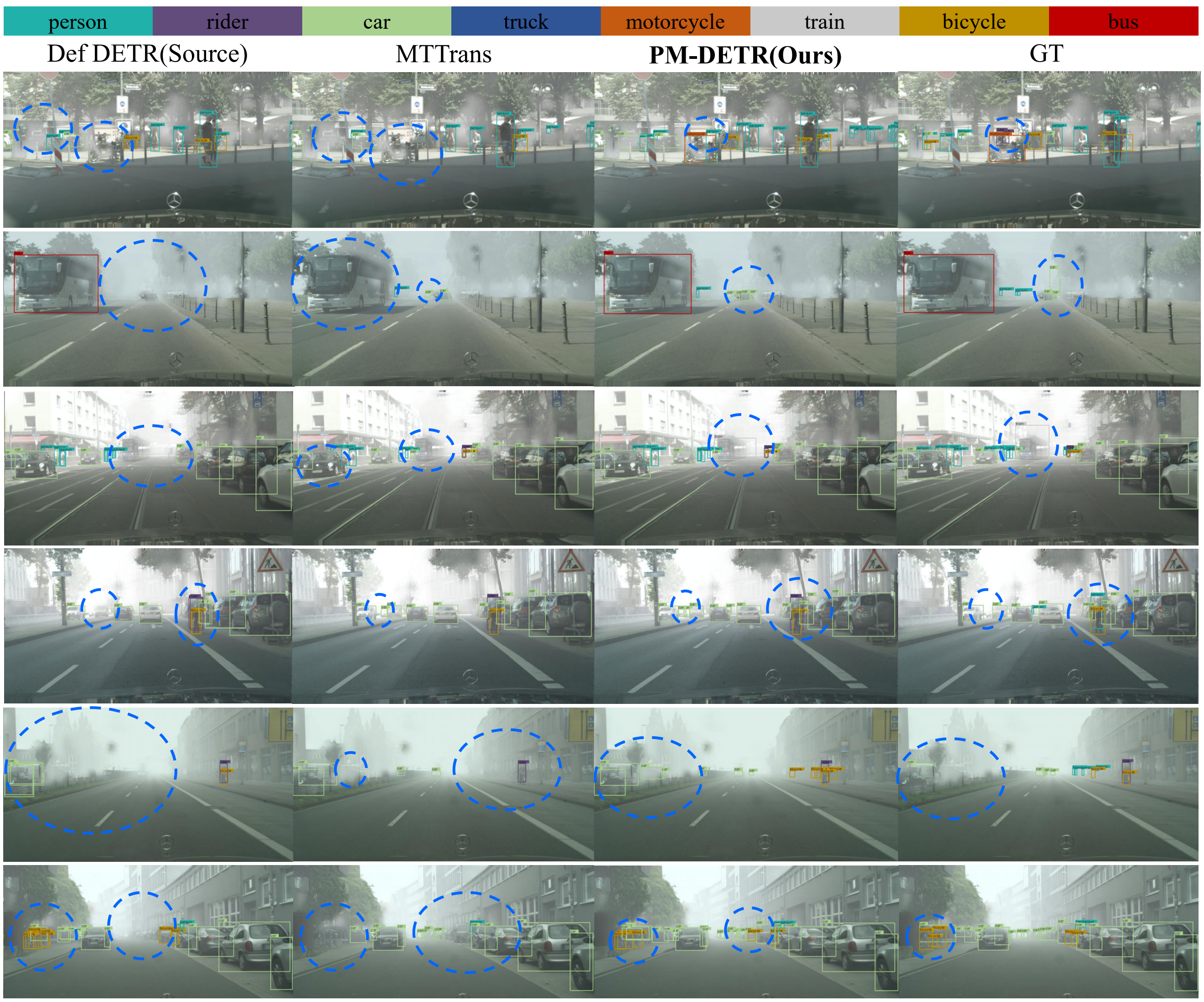


**Figure 1: We compare the performance of different methods on the weather adaptation task, include Deformable DETR as baseline, MTTrans as comparable method, Ours and GT. It can be observed that in the case of Deformable DETR, there are instances of missed detections, false detections, and redundant detections. Meanwhile, MTTrans suffers from the issue of low recall rates, which adversely affects its performance. Our method is even capable of detecting certain errors present in the ground truth annotations.**

distance between two domains, while fully leverage the domain-specific knowledge extracted from the memory space. This will constrain $\hat{d}_{\mathcal{H}_S,\mathcal{H}_T}(S,T)$.

# B ADDITIONAL EXPERIMENTS

## B.1 Details in Dataset Settings

In Cityscapes[4] to Foggy Cityscapes [9] adaptation scenario, all experiments are conducted with a fixed foggy level equals to 0.02. It is important to note that our approach cannot be directly compared to unsupervised learning methods across multiple foggy levels, as doing so would result in a significant increase in training inference time, making it impractical for real-world applications. In Cityscapes to BDD100k[13], we follow the approach outlined in [11] and exclude the "train" category due to its limited availability. Moreover, we assign the following order to the categories based on their ID growth: person, rider, car, bicycle, motorcycle, bus, truck.

In the Mean Teacher framework, the weak augmentation applied to the teacher model includes random horizontal flip, random resize, and random size crop. These transformations help introduce diversity and robustness during training. On the other hand, the strong augmentation employed for the student model includes random Gaussian blur, random grayscale, and color jitter. These additional augmentations further enhance the model's ability to handle variations and improve generalization. For the exponential moving average (EMA) update weights, a value of 0.999 is typically set. This weight helps stabilize the training process and avoids catastrophic forgetting. Following the approach outlined in SFA [11], our discriminative classifier comprises three consecutive multilayer perceptron layers.

## B.2 Comparison With Recent SOTA

As illustrated in Table 1, our method surpasses the performance of all the latest published papers on the challenging weather adaptation scenario.

## B.3 More Visualization Results

In this section, we present additional visualizations that highlight the superior performance of PM-DETR on the Cityscapes to Foggy Cityscapes weather adaptation task. Specifically, we provide visualizations of images from the validation subset of Foggy Cityscapes with the following IDs: [025, 028, 034, 053, 354, 390]. These visualizations demonstrate the effectiveness of our approach in handling the challenges posed by weather variations and showcase the improved results achieved by PM-DETR.

# REFERENCES


[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79, 1 (2010), 151–175.

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. *Advances in neural information processing systems* 19 (2006).

[3] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. 2023. Contrastive Mean Teacher for Domain Adaptive Object Detectors. arXiv:2305.03034 [cs.CV]

[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 *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.

[5] 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.

[6] Jinlong Li, Runsheng Xu, Jin Ma, Qin Zou, Jiaqi Ma, and Hongkai Yu. 2023. Domain Adaptive Object Detection for Autonomous Driving under Foggy Weather. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 612–622.

[7] Wuyang Li, Xinyu Liu, and Yixuan Yuan. 2023. SIGMA++: Improved Semantic-complete Graph Matching for Domain Adaptive Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023), 1–18. https://doi.org/10.1109/TPAMI.2023.3235367

[8] Giulio Mattolin, Luca Zanella, Elisa Ricci, and Yiming Wang. 2023. ConfMix: Unsupervised Domain Adaptation for Object Detection via Confidence-Based Mixing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 423–433.

[9] 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.

[10] Kunyang Sun, Wei Lin, Haoqin Shi, Yu Liu, Zhengming Zhang, Yongming Huang, and Horst Bischof. 2023. AIRA-DA: Adversarial Image Reconstruction Alignments for Unsupervised Domain Adaptive Object Detection. *IEEE Robotics and Automation Letters* 8, 6 (2023), 3645–3652. https://doi.org/10.1109/LRA.2023.3267692

[11] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. 2021. Exploring Sequence Feature Alignment for Domain Adaptive Detection Transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1730–1738.

[12] Xin Yang, Michael Bi Mi, Yuan Yuan, Xin Wang, and Robby T Tan. 2022. Object Detection in Foggy Scenes by Embedding Depth and Reconstruction into Domain Adaptation. In *Proceedings of the Asian Conference on Computer Vision*. 1093–1108.

[13] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. 2018. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687* 2, 5 (2018), 6.

[14] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. 2022. MT-Trans: Cross-domain Object Detection with Mean Teacher Transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer, 629–645.

[15] Jingyi Zhang, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, Xiaoqin Zhang, and Shijian Lu. 2023. DA-DETR: Domain Adaptive Detection Transformer with Information Fusion. arXiv:2103.17084 [cs.CV]

[16] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.