MIC-AQT: Improving Domain Adaptive Object Detection of Adversarial Query Transformers with Masked Image Consistency

Peter Stratton[1] and Xiaoxiao Du[1]

I. Motivation

Deep learning-based object detectors typically suffer from a performance drop when a domain gap (e.g., distribution mismatch) is present between training and testing environments. Unsupervised Domain Adaptation (UDA) methods [1] have been developed to adapt the networks to the target domain by minimizing the cross-domain discrepancy.

Adversarial Query Transformers (AQT) [2] is a state-of-the-art transformer-based domain adaptive object detector. AQT integrates multi-level adversarial feature alignments into a detection transformer (e.g., Deformable DETR [3]) and uses cross-attention modules to classify domain labels and identify domain-specific object features. In this work, we propose to improve the performance of AQT by adding Masked Image Consistency (MIC) [4], a UDA module to improve the learning of target domain context relations. MIC uses a student-teacher network to learn context clues on the target domain by passing masked target images to the student network and unmasked target images to the teacher network. We believe that the added MIC module will encourage the AQT detector to better use context information from pixels close to the objects and will increase the domain adaptability of AQT, particularly for objects that look visually similar to backgrounds. Quantitative evaluation and visual results will be presented to show the domain adaptive object detection performance of our proposed MIC-AQT method.

II. Problem Statement

The UDA problem takes labeled data from the source domain $S = \{x_i^S, y_i^S\}_{i=0}^{N_S}$ and unlabeled data from the target domain $T = \{x_j^T\}_{j=0}^{N_T}$ as inputs to train a neural network f_θ. N_S and N_T are the number of images in the source and target datasets, respectively. The loss functions for UDA networks typically consist of two terms, a supervised source loss L^S and an unsupervised target loss L^T [4]. We also add a masked loss $L^M = H(\hat{y}, p^T)$, where the outputs of the teacher network p^T are used as pseudo-labels for the student network's predictions on the masked image $\hat{y}$. For our detector, L^S and L^M are both implemented with the Hungarian loss [3]. Our loss equation is given by

$$\min_\theta \frac{1}{N_S}\sum_{i=0}^{N_S} L_i^S + \frac{1}{N_T}\sum_{j=0}^{N_T}(\lambda^T L_j^T + \lambda^M L_j^M). \quad (1)$$

$\lambda^{\cdot}$ are weighting parameters ($= 1$ currently). After training, the network is evaluated on the target domain dataset T.

*This work was supported by a grant from Ford Motor Company via the Ford-UM Alliance under award N028603 and the National Science Foundation under Grant IIS-2153171-CRII: III: Explainable Multi-Source Data Integration with Uncertainty.

[1]Department of Robotics, University of Michigan, Ann Arbor, USA {pstratt,xiaodu}@umich.edu

Method	Source	Target	mAP ↑
Source Only (Deformable DETR)	C	FC	27.9
AQT	C	FC/B	43.0/22.9
MIC-AQT	C	FC/B	**44.4/27.0**

TABLE I

Experiment results of MIC-AQT vs AQT. C = Cityscapes, FC = Foggy Cityscapes, B = BDD100K Daytime

III. Results

In all our experiments, we trained using a batch size of 4 on 1 GPU, for 50 epochs on *Foggy Cityscapes* and 20 epochs on *BDD100K*. Table I shows improved detection accuracy (mAP) of MIC-AQT compared to vanilla AQT without the MIC module on both the cityscapes-to-foggy cityscapes [5], [6] and cityscapes-to-BDD100K daytime datasets [7]. Figure 1 shows visual comparison of both methods. As shown, MIC-AQT (blue) correctly identified the cars on the left that blended with the background, while AQT (orange) missed. This shows that MIC-AQT was better at leveraging contextual features to make accurate predictions, even across domains. We also note that MIC can easily work with other transformer-based object detection network architectures (such as AQT) and help improve its performance on domain adaptive object detection tasks. AQT (blue) correctly identified the cars on the left that blended with the background, while AQT (orange) missed. This shows that MIC-AQT was better at leveraging contextual features to make accurate predictions, even across domains. We also note that MIC can easily work with other transformer-based object detection network architectures (such as AQT) and help improve its performance on domain adaptive object detection tasks. AQT (blue) correctly identified the cars on the left that blended with the background, while AQT (orange) missed. This shows that MIC-AQT was better at leveraging contextual features to make accurate predictions, even across domains. We also note that MIC can easily work with other transformer-based object detection network architectures (such as AQT) and help improve its performance on domain adaptive object detection tasks.

References

[1] K. Shen, R. M. Jones, A. Kumar, S. M. Xie, J. Z. HaoChen, T. Ma, and P. Liang, "Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation," in *ICML*, 2022, pp. 19 847–19 878.

[2] W.-J. Huang, Y.-L. Lu, S.-Y. Lin, Y. Xie, and Y.-Y. Lin, "Aqt: Adversarial query transformers for domain adaptive object detection," in *Proc. 31st Int. Joint Conf. Art. Intell.*, 2022.

[3] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," 2021.

[4] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "Mic: Masked image consistency for context-enhanced domain adaptation," in *CVPR*, 2023, pp. 11 721–11 732.

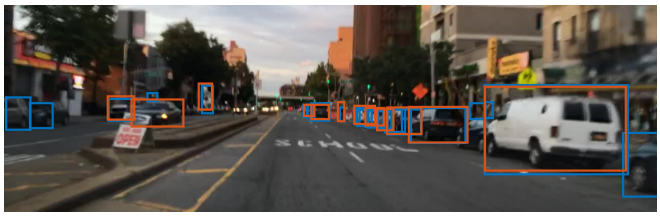

Fig. 1. MIC-AQT Result on BDD100K. Orange are AQT predictions and Blue are MIC-AQT predictions.

[5] 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 *CVPR*, 2016.

[6] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Computer Vision*, vol. 126, pp. 973–992, 2018.

[7] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *CVPR*, 2020, pp. 2636–2645.