Supplementary Materials for MTTrans: Cross-Domain Object Detection with Mean Teacher Transformer

Jinze Yu[1], Jiaming Liu[2], Xiaobao Wei[2], Haoyi Zhou[1], Yohei Nakata[3], Denis Gudovskiy[3], Tomoyuki Okuno[3], Jianxin Li[1], Kurt Keutzer[4], and Shanghang Zhang[2⋆]

[1]Beihang University [2]Peking University
[3]Panasonic Holdings Corporation [4]University of California, Berkeley
yujinze@buaa.edu.cn, shanghang@pku.edu.cn

In this appendix, we will present more implementation details, more ablation studies, and visualization results in other domain adaptation scenarios, such as synthetic to real adaptation and scene adaptation.

A More Implementation Details

Our implementations, including hyper-parameters and model configurations, are based on Deformable DETR [7] and SFA [5] for fair comparisons. Three feature maps (f_1,f_2,f_3) are extracted from the last three layers of the ImageNet [1] pretrained ResNet50 [2] backbone: f_1, f_2, f_3 with a channel dimension of 256, 512, 1024. These feature maps are projected into four feature maps with fixed dimension of 256 by different convolutional layers:

$$z_1 = \text{conv}_1(f_1), z_2 = \text{conv}_2(f_2), z_3 = \text{conv}_3(f_3), z_4 = \text{conv}_4(\text{conv}_3(f_3)) \quad (1)$$

where conv_1 to conv_3 are 1×1 convolutions, and conv_4 is a 3×3, strided 2 convolution. The final input of the encoder z is the concatenated and flattened z_1, z_2, z_3, z_4:

$$z = [z1'; z2'; z3'; z4'] \quad (2)$$

where z_i' is the flattened version of z_i. The dimensions of the encoder and decoder's input and output are 256. We adopt the deep supervision mechanism [3], in which intermediate prediction will be made based on features extracted from some layers of the model. Intermediate features are extracted after each Transformer block of the encoder and the decoder. All feature alignment techniques, including DQFA (domain query-based feature alignment) and TIFA (token-wise image feature alignment) for the encoder, together with DQFA and BGPA (bi-level graph-based prototype alignment) for the decoder, are applied to the corresponding intermediate features. The object detection predictions are made by the decoder's intermediate features as well. We follow FixMatch [4] and Soft Teacher [6] to use different data augmentation for pseudo-label generation, labeled source domain image training, and unlabeled target domain image training. The detailed data augmentation techniques are summarized in Table 1.

⋆ Corresponding Author

Table 1. The details of data augmentation techniques used in this work. We use stronger augmentation for unlabeled target domain prediction and weaker augmentation for target domain pseudo-label generation. "✓" indicates the augmentation is used.

Augmentation	Labeled source training	Pseudo-label generation	Unlabeled target training
Random flip	✓	✓	✓
Random crop	✓	✓	✓
Random resize	✓	✓	✓
Brightness jitter			✓
Constrast jitter			✓
Saturation jitter			✓
Hue jitter			✓
Random Grayscale			✓
Gaussian blur			✓

The number of prototypes in BGPA, and the pre-defined threshold for pseudo-label generation are set by experiment results conducted in the weather adaptation scenario, which will be introduced in Section B later on.

B More Ablation Study Results

B.1 Number of Prototypes in BGPA

As described in Section 3.2, M prototypes are constructed by the BGPA alignment technique. We set M to 9 based on experiments conducted in the weather adaptation scenario, as shown in Table 2.

Table 2. Experiments on the number of prototypes in BGPA conducted in the weather adaptation scenario. Based on the results, the final number is set to 9 in our proposed MTTrans.

Number of prototypes	mAP50
1	43.061
2	43.030
4	43.185
9 (MTTrans)	**43.413**

B.2 Pre-defined Threshold for Pseudo-Label Generation

As described in Section 3.1 and 4.1, the teacher model's predictions are filtered by a pre-defined threshold, and the threshold is set to 0.5 for both classification and bounding box regression tasks based on experiments conducted in the weather adaptation scenario, as shown in Table 3. We tried to set a higher threshold for bounding box regression but obtained inferior results.

Table 3. Experiments on the predefined threshold for pseudo label filtering conducted in the weather adaptation scenario. Based on the results, the final threshold is set to 0.5 for both classification and bounding box regression within the object detection task.

	Threshold for classification pseudo labels	Threshold for bbox regression pseudo labels	mAP50
1(MTTrans)	**0.5**	**0.5**	**43.413**
2	0.5	0.7	42.601
3	0.4	0.5	42.879
4	0.4	0.6	42.664

Note that the threshold, 0.5, is lower than the common threshold in the teacher-student framework for cross-domain object detection based on two-stage detectors, which is usually set to 0.7. We argue that only regions with a high probability of fore-ground categories are retrieved by the region proposal network (RPN) in two-stage object detectors, while a fixed number of object proposals are generated by Deformable DETR. We tried to set the threshold to 0.7 for both tasks in our early attempts, only a small number of objects are kept, and thus the student model can not get enough pseudo labels.

B.3 Ablation Study in Other Domain Adaptation Scenarios

Table 4. Ablation studies in the synthetic to real adaptation scenario, with Sim10k to Cityscapes. MT stands for mean teacher framework, and SharedQE denotes the shared object queries of decoder inputs. DQFA, TIFA, and BGPA represent domain query-based feature alignment, token-wise image feature alignment, and bi-level graph-based feature alignment. Components of other models that differ from MTTrans are marked in red.

Methods	MT	SharedQE	DQFA		TIFA		BGPA		mAP50
			enc	dec	enc	dec	enc	dec	
Deformable DETR (Source)	✗	✗	✗	✗	✗	✗	✗	✗	47.4
MTTrans-AS0(MT-DefDETR)	✓	✓	✗	✗	✗	✗	✗	✗	51.829
MTTrans-AS11	✓	✓	✗	✓	✓	✗	✗	✓	56.812
MTTrans-AS12	✓	✓	✓	✗	✓	✗	✗	✓	56.904
MTTrans-AS13	✓	✓	✓	✓	✗	✗	✗	✓	56.820
MTTrans-AS14	✓	✓	✓	✓	✓	✗	✗	✗	56.863
MTTrans-AS15	✓	✗	✓	✓	✓	✗	✗	✓	56.515
MTTrans-AS21	✓	✓	✓	✓	✓	✓	✗	✗	56.381
MTTrans-AS22	✓	✓	✓	✓	✗	✗	✓	✓	57.330
MTTrans	✓	✓	✓	✓	✓	✗	✗	✓	57.940

Experiments in Section 4.3 are further carried out in the synthetic to real adaptation scenario, and the results are shown in Table 4. From the results, we can observe that the conclusions in the weather adaptation scenario are kept roughly the same: (1) Adding the mean teacher framework directly to Deformable DETR (MTTrans-AS0, MT-DefDETR)) can improve its performance on the target domain (+4.43 mAP). We can notice that it is critical to introduce mean teacher in cross-domain adaptation, but there is still much space for improvement due to the poor quality of pseudo labels. (2) Removing any

aspect of MTTrans (MTTrans-AS11 to MTTrans-AS14)will result in a performance degradation; (3) Altering the alignment technique for the decoder from BGPA to TIFA (MTTrans-AS21, -1.559 mAP), or replacing TIFA for the encoder with BGPA (MTTrans-AS22, -0.610 mAP) both result in performance drop; (4) Removing the shared object queries between the teacher and student models (MTTrans-AS15) will also decrease MTTrans's performance (-1.425 mAP).

C More Visualization Results.

C.1 Pseudo Label Visualization in Other Domain Adaptation Scenarios

More visualization results of the generated pseudo labels and student model's prediction produced in the synthetic to real adaptation and scene adaptation scenarios, as shown in Fig. 1. It can be seen that pseudo labels generated by MTTrans consistently outperform that of the mean teacher version of Deformable DETR (MT-DefDETR), and the pseudo labels are of higher quality compared with students' predictions. However, when viewing the results, we discover some annotation errors in BDD100K, as shown in Fig. 2, which can explain some detection errors, as the City2BDD result shown in Fig. 1

C.2 Detection Results in Other Domain Adaptation Scenarios

We show some visualization results of MTTrans in the synthetic to real adaptation and scene adaptation scenarios, in Fig. 3, and our proposed MTTrans consistently outperforms the baseline models.

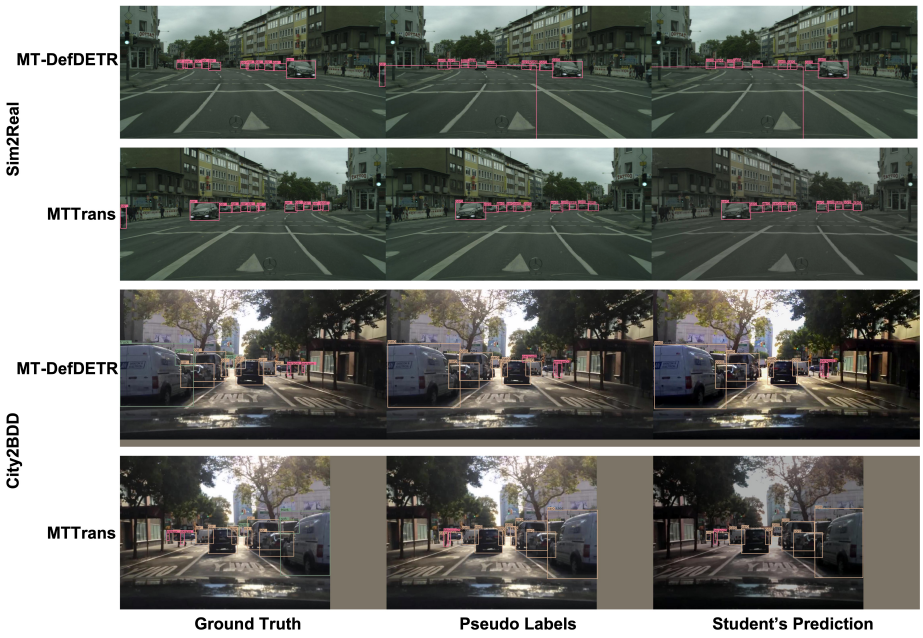

Fig. 1. Visualization results of the generated pseudo labels in the synthetic to real adaptation (Sim2Real) and scene adaptation (City2BDD) scenarios, ground truth annotations, and student model predictions. As can be seen in the visualization result, MTTrans can generate pseudo labels of higher quality compared with MT-DefDETR; and the teacher model performs better than the student model. MT-DefDETR stands for directly applying the mean teacher framework to Deformable DETR.

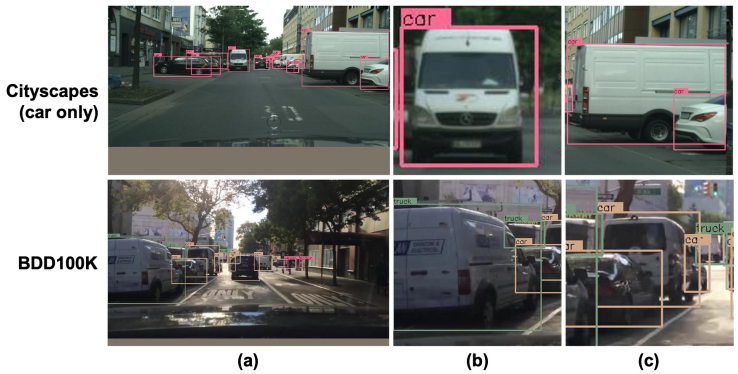

Fig. 2. Annotation errors in BDD100K shown in Fig. 1. Similar "vans" are labeled as cars or trucks in BDD100K, while in Cityscapes, they are labeled as cars.

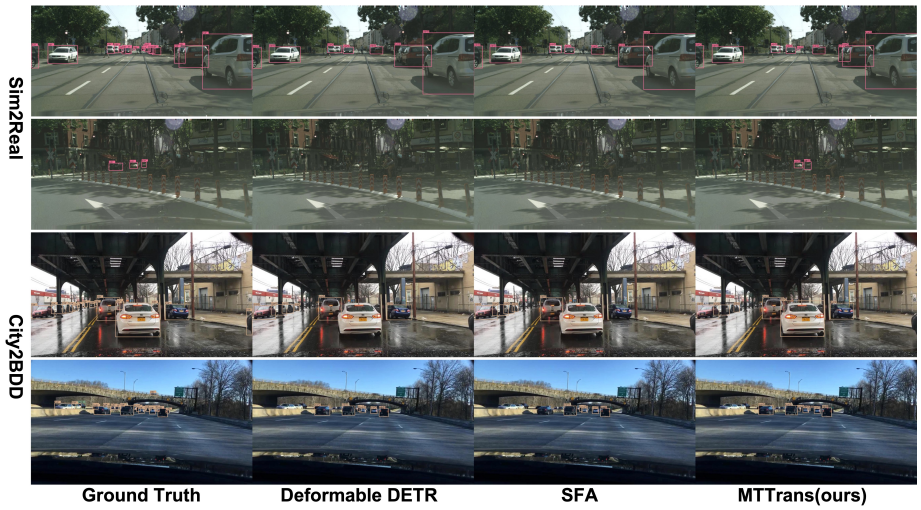

Fig. 3. Visualization of detection results in the synthetic to real adaptation (Sim2Real) and scene adaptation (City2BDD) scenarios. From left to right are ground truth, results obtained by Deformable DETR, SFA, and MTTrans. The predicted category and prediction confidence can be seen on the bounding box labels. Recommend to read with computers, and the original image files will be attached with supplement materials.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
3. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial intelligence and statistics. pp. 562–570. PMLR (2015)
4. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems **33**, 596–608 (2020)
5. Wang, W., Cao, Y., Zhang, J., He, F., Zha, Z.J., Wen, Y., Tao, D.: Exploring sequence feature alignment for domain adaptive detection transformers. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1730–1738 (2021)
6. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3060–3069 (2021)
7. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)