Multi-Domain Knowledge Distillation for Robust Human Object Detection in Diverse Environments

Xiyana Figuera*, Jiun Jeong*, Eunchong Kim*, Isu Jeong* and Soogeun Park*

***Abstract*— In the field of Computer Vision, object detection is a prominent focus of research with groundbreaking success in few-shot learning. Despite progress, challenges still exist in handling diverse real-world scenes. This work focuses on knowledge distillation proposing a shift from compression-focused applications to training models across multiple domains. We introduce a multi-teacher Knowledge Distillation approach for human object detection, addressing the challenges of varied image domains. Our contributions include a multi-domain multi-teacher distillation framework and a novel distillation loss for object detection.**

I. INTRODUCTION

Object detection remains a key focus in current Computer Vision (CV) research [1], with recent achievements, such as the groundbreaking success of few-shot learning of Grounding Dino [2]. Despite these advancements in the area, challenges are prevalent within the field. A notable complexity lies in the diverse and varied nature of real-world scenes, influenced by factors like variations in lighting conditions, backgrounds, and perspectives [3]. These diverse domains in images, represent a significant challenge for models aiming to generalize across different cases.

At the same time, methods such as knowledge distillation (KD) are being widely exploited in deep learning (DL) [4]. KD allows to compress models rendering them suitable for real-time applications. Depending on the type of knowledge the student learns from the teacher [4], KD can be divided into divided into: 1) response-based (logits), 2) feature-based (intermediate representations) and 3) relation-based (relationship tensors). Among them response-based KD does not need of homogeneity of architecture between teachers. This offers a way to harshness different frameworks, especially the combination of several baselines, by leveraging multi-teacher knowledge distillation (MTKD). MTKD is a type of KD, typically implemented by averaging the predictions of multiple teachers [4].

KD is commonly used for compression, however, we propose employing KD to train a student model in multiple domains. Instead of averaging teachers predictions, we leverage the capabilities of convolutional neural networks (CNN) to recognize patterns without explicitly classifying the domain.

This work addresses human object detection in images from multiple domains with varying backgrounds, lighting, angles, and human sizes. Overlapping humans in many instances further increase the difficulty of the task. To overcome these challenges, we propose a multi-teacher knowledge distillation approach, where each teacher specializes in a distinct domain. Specifically, we distinguish three domains: indoor, outdoor, and crowded scenes. We use response-based KD because our teachers have different architectures.

Our contributions are the following:

1) We propose a multi domain multi-teacher distillation framework.
2) We propose a new distillation loss for object detection

Ulsan National Institute of Science and Technology, Ulsan, Korea.
*Equal contribution

II. DATASET

We assess our method using the test set for the "2023 Fall Advanced Computer Vision Final Term Project" CodaLab contest organized by the Vision and Learning Lab at Ulsan National Institute of Science and Technology (UNIST).

We identify three different domains in the test dataset which are indoor, outdoor and crowds. For the indoor domain, we collected our own dataset (we refer to as IN) given that a considerable part of the test data is a customized dataset with scenes from UNIST. We recorded videos of our member in different settings at UNIST with scenes resembling those from the test set. We edited the resolution of all videos to be 640x362 and we picked an image every 20 frames to obtain a dataset which we partially labeled using a tool (cvat.ai) to obtain hard labels and we labeled the rest using YOLOv6 (large). Then, we finetuned YOLOv6 with these labels and obtained dark knowledge for training our student.

For the Outdoor dataset we use the EuroCity Person (ECP) dataset because of the high similarity with the outdoor images in the test dataset. To obtain the dark knowledge for the student model we predicted the labels using YOLOv3-pedestrian. In the case of the Crowd domain, which we include to address the challenge of overlapping humans in the images, we use CrowdHuman (CH) dataset with dark knowledge obtained using ProgressiveDETR

We chose YOLOv3-pedestrian and ProgressiveDETR as teacher models because of the easiness for implementation due to limited time.

III. METHODOLOGY

A. Overview

In this work, we address the problem of human object detection for multiple domain images where the images have different backgrounds, lighting, angles, and the sizes of the humans in the pictures differ. Another challenge is that in many cases humans overlaps in the images further increasing

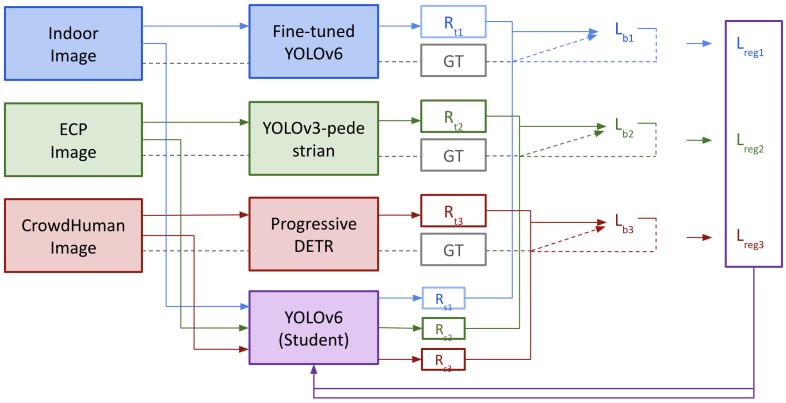

Fig. 1. Multi-domain Knowledge Distillation framework.

the complexity of the task. To address these challenges, we propose the implementation of a multi-teacher knowledge distillation framework (fig. 1) where each teacher is an expert in a specific and different domain (Inddor, Outdoor and Crowded).

In this work, we adopt YOLOv6 as our student model for fine-tuning, specifically we use the small version. YOLOv6 does not only allow us to deploy our method in real-time but it also employs knowledge distillation facilitating a compatible and convenient implementation for knowledge distillation; This is not the case for other members of the YOLO family such as YOLOv7 or YOLOv8.

Given the noticeable differences in patterns between the indoor and outdoor domain, we deviate from common practices of using the average predictions of teachers. We argue that averaging predictions could be detrimental for learning when targeting multiple domains. We posit that the efficiency of the convolutional neural networks in distinguishing patterns might imply that the student can itself discern the domain and learn from each domain without the explicit need for a classification head. In the case of the crowd images, we hope for the model to also discern this domain, however, there are no guarantees for such an assumption.

We also harness the losses that YOLOv6 [5] proposes to employ. We use the varifocal loss for classification and GioU for regression. However, because we have teachers with different architectures, we cannot use the Distribution Focal Loss. This implies that we need to find another method different from the one in YOLOv6 to apply distillation on bounding boxes. This is particularly important because the hidden knowledge from bounding boxes predicted by teachers can mislead the student given that this knowledge is not in the form of soft labels (probabilities) but in the form of regression values. For this reason, we propose to use a bounded loss introduced in [6].

As we can see in equation (1), We combine this bounded loss with the YOLOV6 IoU loss instead of using L1 or L2 to harness GioU loss as IOU losses are found to be mode suitable for bounding box regression [7] than L1/L2 losses to IOU losses. $\mathbf{L}_b$ is the bounded loss, R are the (regression) bounded box information, and s, t and gt denote the student, teacher and ground truth respectively. IoU is the GioU loss. λ and m denote hyperparameters.

$$\mathcal{L}_b(R_s, R_t, R_{gt}) = \begin{cases} \lambda \text{ IoU}(R_s, R_t), & \text{if } a > b \\ 0, & \text{otherwise} \end{cases} \\ \text{where} \quad a = \lambda \text{ IoU}(R_s, R_t) + m \\ b = \lambda \text{ IoU}(R_t, R_{gt}) \tag{1}$$

Equation (2) depicts the loss for fine-tuning our student model. Where CLS is the classification (Varifocal) loss.

$$\mathcal{L}_{reg} = \lambda_c CLS(C_s, C_t) + \lambda IoU(R_s, R_{gt}) + v\mathcal{L}_b \tag{2}$$

We can observe the fine-tuning pipeline of our method in Fig 2.

IV. EXPERIMENTS

In this section we present the results of a comprehensive evaluation of our proposed method. We first compare it with a baseline method of YOLOv6 and further conduct experiments to analyze the contribution to learning of our chosen datasets. Additionally, we assessed is suitable for real-time applications.

A. Baseline

In this experiment, we compared the performance of our approach against a baseline model and a finetuned baseline model. The baseline model we used is the same model we finetuned for our method.

TABLE I
EVALUATION SCORES (MAP) OBTAINED FROM CODALAB

Method	mAP (%)
Baseline	0.172
Fine-tuned baseline	0.224
Ours	0.261

As we can see in table I, our method not only surpassed the baseline but also showed a better performance than the fine-tuned baseline.

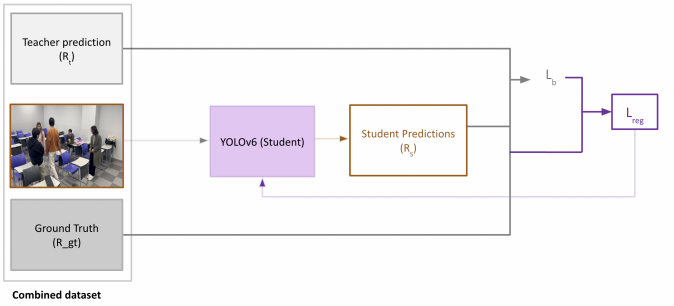

Fig. 2. Fine-tuning pipeline for our student model with our proposed loss.

B. Dataset analysis

We conducted further experiments to understand the impact of individual datasets on the overall performance. We experimented with different combinations to evaluate their positive or negative contributions to the performance of our method.

TABLE II
TEST ACCURACY AFTER TRAINING SECOND TASK USING FEATURE EXTRACTION METHOD

Datasets	mAP (%)
ECP+CH	0.228
IN+CH	0.259
IN+ECP	0.232

In table II, we can notice that the ECP did not contribute significantly to the learning since the combination of indoor and Crowd data achieved a performance almost similar to that of the the full combined dataset.

V. DISCUSSION

In our experimental section we observed that our method improved the baseline achieving an mAP of 0.261 in the test data. This demonstrates that this method has the potential to be integrated in the fine-tuning stage of other methods such as YOLOv6. Further experiments, specially ablation studies are required to fully determine whether the multi-domain learning, the proposed loss or the combination of both can be integrated to existing pipelines.

The effective selection of datasets is also a crucial aspect for our method. In our experiments related to our datasets, we learned that the choice of the ECP dataset did not have a significant impact in improving the performance of the baseline. A possible reason for this is that the teacher model we chose (YOLOv3 Pedestrian) is not the baseline in the ECP dataset; it ranked 7th.

One more essential aspect is the suitability of our method for real-time application. For our method, we observed efficient performance as the model successfully processed 448x448 color images on a single RTX3090 GPU, completing the task in just 4.6149 milliseconds and achieving a frame rate of 209.23 frames per second.

VI. IMPLEMENTATION DETAILS

A. Intructions on how to run the code

We kept the same workflow as YOLOv6, for this reason, the instructions to run the code from YOLOv6 are very similar to ours. Detailed instruction is written in our readme file. The importan difference is that we defined our loss as MTKD and to finetune using our method it is necessary to pass the argument –MTKD. (Please refer to read me for command line to finetune)

Inference can be run with the normal command of YOLOv6 by simply specifying the correct path for the finetuned model.

B. Implementation details

For our implementation we added a new loss to a file called `loss_teacher_bbox.py`. We use the `distil_loss.py` available in YOLOv6 as the base and implemented the losses in eq(1) and eq(2).

We also created a new files for the dataloader called `dataset_with_gt.py` and `data_load_gt.py` these files have two different targets, one for the teachers dark knowledge and one for the our ground truth.

We also edited `engine.py` to include our loss and the new dataloader. And we also edited train to include the argument MTKD so we can use our method by only passing this argument.

More details such as complete paths are available in the readme file. Codes created or edited for our implementation have comments that start with ### Ours, please refer to them for more specific information of the implementation.

VII. CONCLUSION

In this work, we propose a multi-teacher knowledge distillation fine-tuning method to improve the robustness of models under variations of light, object sizes and backgrounds.

We propose the use of MTKD as a way to learn multiple domains harnessing knowledge from teachers expert in a specific domain, without the restriction of a homogeneous architecture. We also propose a new loss by combining the a bounded regression loss with the loss of YOLOv6.

Our model surpasses the baseline, even after the baseline is finutuned. This results are promising and we plan to conduct more experiments in the future to further evaluate and validate our findings.

REFERENCES

[1] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023.

[2] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[3] M. Ahmed, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Survey and performance analysis of deep learning based object detection in challenging environments," *Sensors*, vol. 21, no. 15, p. 5116, 2021.

[4] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[5] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[6] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," *Advances in neural information processing systems*, vol. 30, 2017.

[7] G. Z. S. Loss, "More powerful learning for bounding box regression [j]," *arXiv preprint arXiv:2205.12740*, 2022.

Team member name	Role and responsibility	Contribution score
Isu Jeong	Indoor data collection, Idea proposal, Teacher model Implementation, Final presentation preparation	20%
Jiun Jeong	Indoor data collection, Outdoor & Crowd dataset collection, Indoor data labeling, Final presentation preparation	20%
Xiyana Figuera	Indoor data collection, MTKD Idea proposal, MTKD and bounding box distillation loss implementation, codes for combining datasets, Mid presentation, MTKD additional experiments, MTKD results analysis	20%
Soogeun Park	Indoor data collection, indoor data processing, Dataset format conversion, Mid presentation preparation	20%
Eunchong Kim	Indoor data collection, Teacher model Implementation, MTKD additional experiments, MTKD results analysis, Mid presentation preparation, Final presentation	20%

Jiun Jeong submitted the codes and dataset for our team.