ISBN: 978-981-18-7950-0
2023 the 13th International Workshop on Computer Science and Engineering (WCSE 2023)
doi: 10.18178/wcse.2023.06.023

Communication-efficient Multi-source Domain Adaptive Object Detection under Privacy Constraints

Peggy Joy Lu [1,2+] and Jen-Hui Chuang [1]

[1] Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[2] National Center for High-Performance Computing (NCHC), Taichung, Taiwan

Abstract. To establish a more generalized model in object detection, collaboration among multiple cameras can increase data diversity and reduce the effort for data collection, leading to a new research area as multi-source domain adaptative object detection (MSDAOD). However, preserving source data privacy in MSDAOD is challenging due to the lack of information integrated from all source domains. In this paper, we present an architecture that allows multiple clients protect the privacy of their own local data while the server only access target data. First, we analyze the effectiveness of using multiple sources, in domain adaptive object detection task. In client sides, we propose a source-only probabilistic teacher (PT) and leverage probabilistic teacher for domain adaptation (PTDA) as detectors to reduce false negatives. Moreover, we also introduce a Pseudo-label Voting Mechanism to filter out false positives with minimal communication costs. The performance of the proposed approach is evaluated on the ck2b and skf2c datasets and compared with other multi-source domain adaptation as well as federated learning methods. To sum up, the proposed method achieved better performance while preserving source data privacy and minimizing communication costs, without requiring the same model structure among different clients.

Keywords: Domain adaptive object detection, multi-source domain adaptation, federated learning, source-free domain adaptation, privacy preservation

1. Introduction

In surveillance systems, object detection is widely employed in many practical applications. To obtain a domain generalization model from varied images obtained in different scenes (domains), a deep learning-based task can be collaboratively trained by different edge devices, i.e., cameras. However, due to the data privacy concern, edge devices may not be able to share their local data with others. Such situation can be considered as a source-free multi-source domain adaptation problem in object detection.

A number of unsupervised domain adaptive object detection methods have been proposed in recent years, with some adopting domain adversarial training technique [1] while others using pseudo label-based self-training [2] or mean teacher training [3]. Recently, there has been an emergence of multi-source domain adaptation for object detection approaches, such as DMSN [4], MTK [5] and TRKP [6]. Despite improved accuracy achieved by some of these methods, data privacy concerns have not been adequately addressed, as they require the access to both the multi-domain source data, as well as the target data, during training. For data privacy consideration, source-free domain adaptation (SFDA) [7] tackles the problem of domain adaptation with unlabelled target data available during the training stage, but only for a single source domain adaptation. While FedPT [8] proposed a scenario also for source-free multi-source domain adaptation in object detection by leveraging federated learning architecture, the approach only considers source data privacy, with target data accessible to all edge devices. For model aggregation in federated learning, consistent model structures across different clients and back-and-forth model transmission are typically required, while the former lacks flexibility and the latter is often time-consuming.

+ Corresponding author. Tel.: + 886-4-24620202#850.
E-mail address: peggylu@narlabs.org.tw.

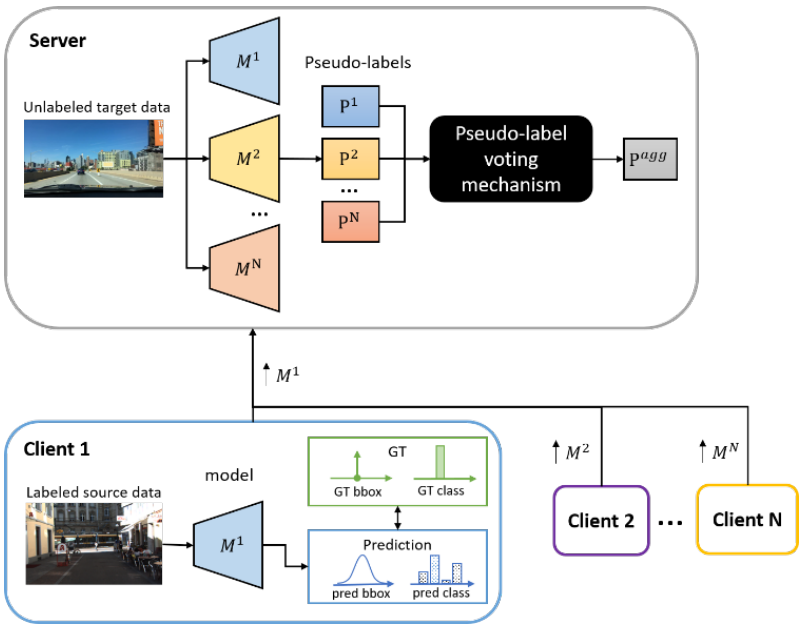


Fig. 1: Architecture of our system. Clients train models on local source data and send those models back to the server. The server generates pseudo-labels on target data from different models and aggregate them by *Pseudo-label Voting Mechanism* to obtain ensembled pseudo-labels as the detection results.

In this paper, we adopt the architecture of multi-source domain adaptive object detection presented in FedPT [8], as shown in Fig. 1, with the communication cost reduced by eliminating model transmission. First, we classify pseudo-labels generated from different models by the ground-truth and analyze the effectiveness of using multiple sources. In client sides, we leverage the technique of Probabilistic Teacher (PT) [9] and convert it to a source-only version. Furthermore, we try to reduce false negatives (FNs) by adopting the source-only PT, as well as PT for Domain Adaptation (PTDA). Finally, we propose a *Pseudo-label Voting Mechanism* to further filter out false positives (FPs) and enhance the detection results.

Thus, contributions of the paper include:

- Classifying pseudo-labels generated from different models, and analyzing the effectiveness of using multiple sources, in domain adaptive object detection
- Proposing *source-only PT* on clients to reduce FNs.
- Proposing a *Pseudo-label Voting Mechanism* to filter out FPs and improve the detection results.

2. Proposed approaches

2.1. System Architecture

Fig. 1 presents an overview of our framework, which involves N clients that train a model M^n on their private local data.[1] To improve the detection performance, clients leverage the PT technique to establish uncertainty on both classification and bounding box (bbox) regression. After the training, each client will send the model back to the server. The server will then collect pseudo-labels generated from different models and filter out FPs using the *Pseudo-label Voting Mechanism*. The resulting ensembled pseudo-labels are considered the final detection results without additional training.

2.2. Pseudo-label classification in multi-source domain adaptation

To evaluate the potential benefits of incorporating multi-source data in domain adaptation for improving the detection results, we first combine the detection results from multiple source domains and classify them based on their overlap conditions, as shown in Fig. 2. Assuming there are two sources (S_A and S_B), and the models trained on S_A and S_B are denoted by M_{SA} and M_{SB}, respectively. In addition, pseudo-labels obtained by

[1] Note that the models used by different clients can have different architectures and backbones, such as YOLO [21] or Faster R-CNN [16], and VGG16 [15] or ResNet50 [22].

applying these models on the target domain are represented as P_{SA} and P_{SB}, respectively, while GT denotes the ground truth of the target domain.

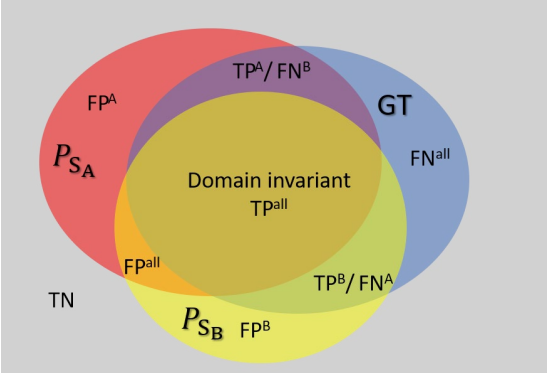


Fig. 2: Classification of different types of detection results from multiple source domains based on their overlap conditions.

Regarding the evaluation metrics, we categorize the detection results according to the overlap conditions of: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Specifically, TP^{all} indicates the object that can be accurately detected by all source models, and can be considered as the domain-invariant part of source and target domains. On the other hand, FN^{all} refers to the scenario where the object cannot be detected by either M_{SA} or M_{SB}. For other false negatives, i.e., FN^B (TP^A) and FN^A (TP^B), only one of the models can detect the object correctly. For example, FN^A indicates that M_{SA} cannot detect the object, while the results correspond to TP^B as M_{SB} can detect it correctly. Similarly, for false positives, FP^A and FP^B indicates that M_{SA} and M_{SB} will falsely detect the object, respectively, while FP^{all} means all models in the source domains detect the object mistakenly.

In the case of FP^{all} and FN^{all}, where all models yield incorrect detection results, incorporating multiple sources cannot help to improve the overall system performance. In addition, for objects detected by only one model, say M_{SA}, it can be difficult to determine whether it belongs to FP or TP, and there is a trade-off between filtering out FP^A and TP^A. Therefore, it is important to reduce FNs in the initial stage, before the system can subsequently filter out FPs (FP^A and FP^B) to achieve the better results of object detection.

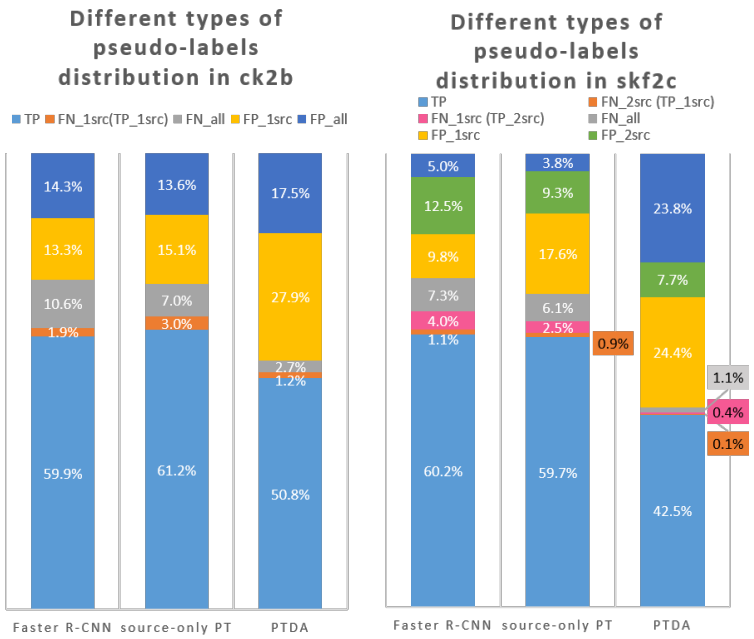
2.3. Reducing false negative by source-only Probabilistic Teacher

To reduce false negatives (FNs), two detectors are used on the client side: Probabilistic Teacher for Domain Adaptation (*PTDA*) [9] and the *source-only PT*. The original PT is designed to solve unsupervised domain adaptation problems, so it requires both source and target data during training. To address this limitation, we propose a source-only PT which only access the source domain data.

In Faster R-CNN [16], uncertainty exists only in the classification task but not in bbox regression, as the former uses cross-entropy while the latter uses L1 loss for training. To establish uncertainty in bbox regression, each bbox coordinate is modelled with a Gaussian distribution. As the proposed source-only PT provides probabilistic estimates of both object classification and bbox regression, such scheme can improve the overall system performance by reducing FNs even without using target data on the clients. In Section 3.2, the effectiveness of both PTDA and the source-only PT in reducing FNs will be demonstrated. However, PTDA may result in a large number of FPs, which need to be further suppressed on the server side, as discussed next.

2.4. Filtering out false positives by Pseudo-label Voting Mechanism

As depicted in Fig. 2, FNs and FPs are categorized based on their overlap conditions. However, in unsupervised domain adaptation, distinguishing TP^A and FP^A for M_{SA} without ground-truth is challenging, and there is a trade-off between filtering out FP^A and/or TP^A, as discussed in Sec. 2.2. Therefore, it is crucial to look into the relative amount of FPs and TPs (FNs) before taking any filtering decisions.



(a) (b)

Fig. 3: The percentage of different result types for different detectors in (a) *ck2b* and (b) *skf2c*.

Fig. 3 (a) illustrates the percentage of different types of detection results in the experiments for dataset *ck2b*, where FN_1src (TP_1src) denotes that only one detector missed the object, while the other correctly detect it, i.e., the sum of FN^A (TP^B) and FN^B (TP^A) in Fig. 2. As for the *skf2c* results depicted in Fig. 3 (b), on the other hand, as there are three sources, FN_2src (TP_1src) indicates that two detectors missed the object while one detected it successfully. After adopting the proposed source-only PT and PTDA, the percentage of TP_1src are extremely low (source-only PT: 3.0% and 0.9% for *ck2b* and *skf2c*, respectively; PTDA: 1.2% and 0.1 % for *ck2b* and *skf2c*, respectively). On the contrary, the percentage of FP_1src increase a lot (source-only PT: 15.1% and 17.6% for *ck2b* and *skf2c*, respectively; PTDA: 27.9% and 24.4% for *ck2b* and *skf2c*, respectively).

Based on the above observations, a *Pseudo-label Voting Mechanism* is proposed, which will

1. Match pseudo-labels generated from different detectors with the largest IoU (intersection over union) bbox, which also has an overlap ratio larger than 0.8.
2. Filter out a bbox that does match any other bbox in the previous step, i.e., the object is only detected by one detector, which may correspond to TP_1src or FP_1src and is regarded as an FP.[2]
3. For each group contains more than one pseudo-label (bbox), average the bbox coordinates, including their uncertainties, to obtain a new bbox.

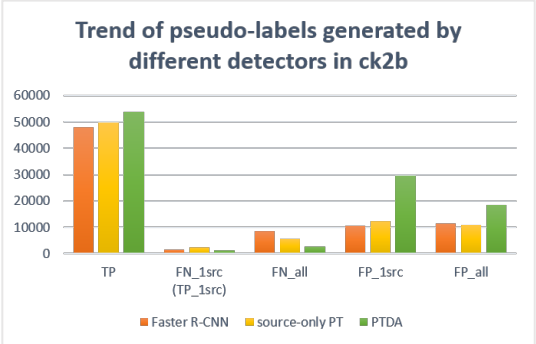
3. Experiments

3.1. Experimental setting

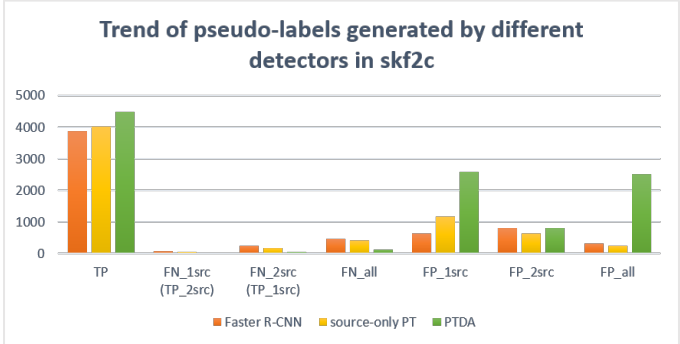
We evaluated the proposed system on two different domain adaptation scenarios, i.e., *ck2b* (source domains: Cityscapes[10] and KITTI [11]; target domain:BDD100k [12]) and *skf2c* (source domains: Sim10k [13], KITTI and Foggy Cityscape [14]; target domain: Cityscape). We implemented the domain adaptive object detection using Detectron2 [17] and employed Faster R-CNN [16] as the detector and VGG16 [15] as the backbone. The remaining parameters following the standard setting proposed in [1].

[2] Since the ratio of FP_1src is much higher than TP_1src, lowering FP_1src is likely to improve the overall system performance, even with a small number of TPs being sacrificed.

3.2. Pseudo-label analysis



(a) ck2b dataset



(b) skf2c dataset

Fig. 4: The trend of pseudo-labels generated by different detectors for different types. The detectors used in the experiments include Faster R-CNN, source-only PT and PTDA.

Fig. 4 demonstrates the trend of different types of detection results, as defined in Sec. 2.4, after adopting selected Probabilistic Teacher techniques. It can be observed that both *source-only PT* and *PTDA* can improve TP detection. In addition, despite the original FNs in Faster R-CNN is already low, PT further reduces them. However, the use of PT also increases FPs, especially for PTDA, which may adversely affect the final precision. Therefore, we employ Pseudo-label Voting Mechanism to further filter out the FPs, as discussed in Sec. 2.4.

3.3. Experimental results

Table 1 shows the experimental results evaluated on the common category, *car*, in terms of the widely used average precision (AP). Details of all proposed and compared state-of-the-art methods are given in the following:

i. Multi-source domain adaptation methods (MSDA): It uses multiple source data and the target data simultaneously, which does not take data privacy into consideration. All clients send their dataset to a central server for training, which may result in a higher communication cost than sending a model since datasets are typically larger than models.

ii. Federated learning (FL): An architecture is designed to collaboratively train a global model while preserving local data privacy. It involves multiple clients and one server, where the server sends a global model for the clients to train. Therefore, the models used by different clients must have consistent structures with the global model. The clients then send back their updated models to the server. This process may repeat for many rounds, and each round requires the transmission of $N+1$ models for N clients. This increases the communication cost significantly, especially when N is large or the model is complicated.

iii. Proposed methods: We propose a *source-only PT* and leverage the *PTDA* as two detectors in the clients, wherein only local data can be accessed in the former while the target data is also available in the latter. Unlike FL, the proposed method only requires the clients to transmit their model to the server once and the model structure in each client in not constrained, resulting in a significantly lower communication cost and flexible detector types.

iv. Oracle: It is an upper bound of system performance, obtained by using fully labeled target in the training process.

One can see from Table 1 that for *ck2b*, TRKP [6] achieves the best performance but it did not consider data privacy, as it requires the access to the complete source datasets. On the other hand, the proposed method outperforms other MSDA and FL methods by more than 3% for *ck2b*. As for *skf2c*, the proposed method uses *PTDA* and *source-only PT* as detectors in the clients and achieves the best (64.1%) and the second-best (59.8%) AP, respectively, while preserving the source data privacy. Although the performance of source-only PT is not as good as PTDA, it only uses source data during training while PTDA uses both source and target data. In the case of Federated Learning, the performance is not good enough even when PT is adopted in the clients (FedPT

[8]), which also incurs additional communication costs between the server and clients. In summary, the two proposed methods outperform most other methods while under more challenging privacy constraint and have lower communication costs. However, there are still some gaps between the proposed methods and Oracle, and need to be investigated further.

Table 1: Results of domain adaptation. The average precision (AP, %) on *car* category in target domain is evaluated.

Method type	Privacy preserving		Communication	methods	AP on car	
	Source	Target			ck2b	skf2c
i. MSDA	×	×	*N* datasets	MDAN [19]	43.2	50.0
				M^3SDA [20]	44.1	50.7
				DMSN [4]	49.2	--
				MTK [5]	--	52.9
				TRKP [6]	**58.2**	--
ii. Federated Learning	✓	×	3 rounds × (*N*+*1*) models	FedPT [8]	48.1	--
	✓	✓	3 rounds × (*N*+*1*) models	FedAvg [18]	43.3	51.1
iii. Proposed method	✓	×	1 round × *N* models	PTDA	52.7	**64.1**
	✓	✓	1 round × *N* models	source-only PT	52.2	59.8
iv. Oracle	×	×	×	Oracle	60.2	66.4

4. Conclusion

In this study, we presented a novel approach for multi-source domain adaptive object detection without revealing source data. We classified the detection results and analyzed the trend of them after adopting probabilistic teacher techniques. Our approach utilizes source-only PT and PTDA as detectors in the clients to reduce false negatives, and a Pseudo-label Voting Mechanism is proposed to further filter out false positives. Through experiments on the popular ck2b and skf2c datasets, we demonstrated that our method is capable of preserving source data privacy, lowering communication costs, providing detector flexibility, and improving detection accuracy. Overall, our approach provides a promising direction for privacy-preserving domain adaptive object detection. Future work can focus on more effectively leveraging of information from multi-source data to generate a domain invariant model.

5. Acknowledgements

This work was supported in part by National Science and Technology Council of Taiwan R.O.C. under NSTC 111-2634-F-A49-010 and NSTC 110-2221-E-A49-083-MY2.

6. 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," *Proceedings of the IEEE conference on CVPR*, 2018, pp. 3339–3348.

[2] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin, "Collaborative training between region proposal localization and classification for domain adaptive object detection," *European Conference on Computer Vision*. Springer, 2020, pp. 86–102.

[3] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan, "Unbiased mean teacher for cross-domain object detection," *Proceedings of the IEEE/CVF conference on CVPR*, 2021, pp. 4091–4101.

[4] Yao, Xingxu and Zhao, Sicheng and Xu, Pengfei and Yang, Jufeng : "Multi-source domain adaptation for object detection", *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[5] Zhang, Dan and Ye, Mao and Liu, Yiguang and Xiong, Lin and Zhou, Lihua : "Multi-source unsupervised domain adaptation for object detection", *Elsevier Information Fusion*, Vol. 78, pp. 138-148, 2022.

[6] Wu, Jiaxi and Chen, Jiaxin and He, Mengzhe and Wang, Yiru and Li, Bo and Ma, Bingqi and Gan, Weihao and Wu, Wei and Wang, Yali and Huang, Di, "Target-Relevant Knowledge Preservation for Multi-Source Domain Adaptive Object Detection", *Proceedings of the IEEE/CVF Conference on CVPR*, pp. 5301-5310, 2022

[7] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang, “A free lunch for unsupervised domain adaptive object detection without source data,” *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 8474–8481.

[8] Peggy Joy Lu, Chia-Yung Jui and Jen-Hui Chuang, “Federated Multi-source Domain Adaptive Object Detection with Probabilistic Teacher”, *International Conference on Industrial Application Engineering (ICIAE)*, 2023.

[9] Chen, Meilin, et al., “Learning domain adaptive object detection with probabilistic teacher”, *arXiv preprint arXiv:2206.06293*, 2022.

[10] Cordts, Marius, et al., “The cityscapes dataset for semantic urban scene understanding”, *Proceedings of the IEEE conference on CVPR*, 2016.

[11] Geiger, Andreas, Philip Lenz, and Raquel Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite”, *Proceedings of the IEEE conference on CVPR*, 2012.

[12] Yu, Fisher, et al.: “Bdd100k: A diverse driving video database with scalable annotation tooling”, *arXiv preprint arXiv*:1805.04687, 2018.

[13] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan, “Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?,” *arXiv preprint arXiv*:1610.01983, 2016.

[14] Christos Sakaridis, Dengxin Dai, and Luc Van Gool, “Semantic foggy scene understanding with synthetic data,” *International Journal of Computer Vision*, 2018.

[15] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv*:1409.1556, 2014.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[17] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick, “Detectron2,” https://github.com/facebookresearch/detectron2, 2019.

[18] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, “Communication-efficient learning of deep networks from decentralized data,” *Artificial intelligence and statistics*. *PMLR*, 2017.

[19] Han Zhao, Shanghang Zhang, Guanhang Wu, Jos´e MF Moura, Joao P Costeira, and Geoffrey J Gordon, “Adversarial multiple source domain adaptation,” *Advances in neural information processing systems*, vol.31, 2018.

[20] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang, “Moment matching for multi-source domain adaptation,” *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[21] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M., “Yolov4: Optimal speed and accuracy of object detection”, *arXiv preprint arXiv:2004.10934*, 2020.

[22] He, K., Zhang, X., Ren, S., & Sun, J, “Deep residual learning for image recognition”, *Proceedings of the IEEE conference on CVPR,* pp. 770-778, 2016.