Cross-domain One-shot Video Object Detection

Yusuf K. Hanoglu[†]
Dept. of Electronics and Comm. Eng.
Istanbul Technical University

Bilge Gunsel[†]
Dept. of Electronics and Comm. Eng.
Istanbul Technical University

Filiz Gurkan[†]
Dept. of Electrical-Electronics Eng.
Istanbul Medeniyet University

***Abstract*—One-shot-object detection (OSOD) aims to detect novel object classes using a single example of an unseen class. Cross-domain OSOD is a more challenging problem since the seen and unseen objects are sampled from the entirely disjoint datasets. The majority of the existing CD-OSOD methods focus on image datasets where the video domain remains largely unaddressed. To tackle this problem, we introduce a one-shot cross-domain video object detection (CD-OSVOD) model enabling adaptation from the still image to the video. Specifically the novel target object is designated as the query shot and a target driven cross-domain finetuning (FT) scheme is integrated with a baseline object detector. To address the requirements of the long term video object detection, the FT scheme is augmented with a novel Online Target Update (OTU) mechanism, enabling the detector to handle challenges such as appearance changes and occlusions. The OTU is controlled by a temporal aggregation module (TAM) which leverages temporal information in video and triggers update of the one-shot query when the temporal consistency is disrupted. The proposed CD-OSVOD utilizes base models trained on COCO and VOC still image datasets and successfully adapts to the video domain for novel object classes. Performance evaluations on challenging VOT-LT benchmarking video dataset demonstrate significant improvement in AP50 and mAP scores, highlighting the effectiveness of the proposed domain adaptation approach.**

***Index Terms*—Video object detection, cross-domain learning.**

I. Introduction

Video object detection aims to localize and classify objects of interest across the video frames. Although recent progress in deep learning led to build object detectors with high accuracy, achieving robust generalization performance requires large annotated datasets, making training data preparation labor-intensive. In addition, generalization to unseen (novel) object classes requires complicated retraining processes. Few shot object detection (FSOD) methods are introduced to transfer the knowledge gained on data-abundant seen (base) classes in the training phase to the data-scarce novel classes in the inference phase. Subsequently , one-shot object detection (OSOD) has emerged as a more challenging approach, aiming to detect all instances of a novel class using only a single query sample of an unseen object.

Early studies on FSOD and OSOD primarily focused on natural still images, assuming that the training and test sets share the same category labels [1]–[3]. More recent OSOD research has extended to video object detection. Several models have been proposed, such as tube proposal network with temporal matching network [4] and self-supervised spatial-temporal feature enhancement for one-shot video object detection [5]. [6] introduces QDETR, which leverages information from the query image along with the spatio-temporal context of the target video, significantly improving the precision of target object localization. Differ from these models, the recent works focus on cross-domain scenarios (CD-OSOD), where the source and target domains contain entirely disjoint object categories. Distill-cdfsod [7] introduces several still image datasets and a distillation-based cross-domain learning method. CD-ViTO [8] employs domain-enhanced vision transformers to improve feature alignment across domains.

However, existing CD-OSOD approaches predominantly restrict domain shifts to still images, overlooking the challenges of adapting models from images to videos. This limits the use of CD-OSOD in real-world video applications, where objects exhibit temporal variations and motion-induced distortions. Additionally, this requires repeat after the base training of the network in the related domain, that consumps an extensive training effort.

To address these challenges, we propose a novel Cross-Domain One-Shot Video Object Detection (CD-OSVOD) framework. Our main contributions are as follows: i) We propose an inference network that facilitates the cross-domain learning through a one-shot finetuning layer and the Temporal Aggregation Module (TAM) integrated into the architecture. To tackle the online processing requirements of various video -based applications, our approach adopts a one-shot learning strategy. ii) We introduce an unsupervised target query shot update mechanism to enhance the cross-domain video object detection. Specifically, in order to improve robustness to occlusion and appearance changes of the query object across the video frames, we augment the one-shot finetuning with a novel Online Target Update (OTU) mechanism controlled by TAM. iii) Differ from the existing work, we leverage the base OSOD models trained on the still image datasets, to alleviate the need for an extensive base training in the video domain. Consequently, the proposed CD-OSVOD framework enables cross-domain adaptation without retraining on the base object classes.

II. Learning The Base Object Classes

In a one-shot video object detection (OSVOD) setup, the training procedure is performed in two-stages. The first stage, referred to as base training, aims to learn generalizable representations from a large-scale dataset D_{base} that contains abundant annotated instances of seen (base) object classes.

[†] Equal contribution.

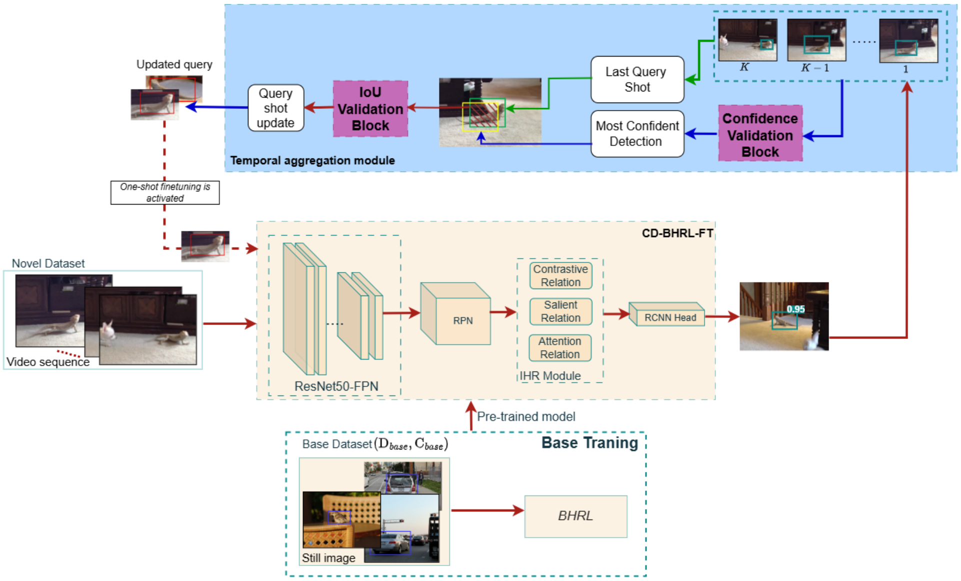

Fig. 1: Overall inference architecture of CD-BHLR-OTU.

At the second stage referred to as the novel training, the domain adaptation on novel set D_{novel} with only one sample for each unseen (novel) class is achieved. It is important to note that in the OSOD setup, the base classes the base classes C_{base} and novel classes C_{novel} are non-overlapping, i.e., $C_{base} \cap C_{novel} = \emptyset$. In addition, the cross-domain OSOD (CD-OSOD) tackles a more realistic scenario in which the distribution of the base domain P_{base} differs from that of the target novel domain P_{novel}. Under this setting, the ultimate goal of the CD-OSOD is to train a robust object detector based on the D_{base}, then localize and classify unlabelled objects of a novel query set D_{query} with object classes C_{query}, where $C_{query} \subset C_{novel}$. For our cross-domain one-shot video object detection (CD-OSVOD) setting, the base dataset consists of still images from either the COCO [9] or VOC [10] datasets. We used the challenging VOT-LT [11] as the novel dataset, due to it is a benchmarking dataset in the long-term video object tracking.

A. Training Network Architecture

In our cross-domain framework, we adopt the Balanced and Hierarchical Relation Learning (BHRL) network [2] as the base training model due to its strong performance on still images. Originally designed as a one-shot object detector without fine-tuning mechanism, BHRL is denoted as BHRL-C in our setup to distinguish it from the designed novel training network. The BHRL-C architecture comprises three main components: a Region Proposal Network (RPN), an Instance-Level Hierarchical Relation (IHR) module for multi-level relation modeling, and an R-CNN head for final detection and classification.

Given a target object query $Q \subset D_{query}$ and an input image I, feature maps are extracted using a ResNet-50 backbone integrated with a Feature Pyramid Network (FPN). To associate the query object with regions in the input image, a similarity map is computed between the feature representations of Q and I using the SiamMask mechanism. This similarity map is then passed to the RPN, which generates a set of candidate bounding box proposals likely to contain the queried object. The IHR module processes the similarity maps of these proposals to learn and integrate attention-level, contrastive-level, and salient-level relations. The output of the IHR module is forwarded to the R-CNN head, which performs the final object detection and classification of instances related to the query.

B. Training Loss

During the training phase of BHRL-C [2] on the base object classes, one-shot parameter learning is performed using a ratio-preserving loss function, as defined in Eq. 1.

$$\mathcal{L}_{base} = \sum_{C_{base} \in D_{base}} (\frac{1}{N}(u \sum_{i \in R_p \cup R_{fp}} \mathcal{L}^i_{CE} + v \sum_{i \in R_{tn}} \mathcal{L}^i_{CE})) \tag{1}$$

In Eq.1, $\mathcal{L}^i_{CE}$ denotes the softmax cross-entropy loss value for the proposal i. R_{fp} and R_{tn} respectively denote the set of false positives and true negatives. N is the number of proposals generated by RPN. BHRL improves the one-shot learning via the dynamic weights u and v formulated in Eq.2 where α is the static sample balancing rate. N_{fp} and N_{tn}, respectively denote the number of false positives and true negatives, and N_p refers the number of positive samples for the given base query from class C_{base}.

$$u = \frac{N \cdot \alpha}{N_p + N_{fp}}, \quad v = \frac{N \cdot (1 - \alpha)}{N_{tn}} \tag{2}$$

III. Detection of The Novel Video Objects

The inference network architecture of the proposed CD-OSVOD system is illustrated in Fig. 1. Unlike the existing works [7], [8], we first perform the base training on still-image object detection datasets, then apply cross-domain adaptation to novel classes in the video dataset. By building on pre-existing "off-the-shelf" base detectors trained on still-image datasets, our method avoids lengthy base-training cycles while still achieving adequate performance in the video domain.

Specifically, the designed detector referred to as CD-BHRL-OTU employs the base models trained by the vanilla BHRL-C on COCO or VOC still image dataset objects and made available to public access [2]. The cross-domain adaptation is achieved by the integration of a one-shot fine-tuning layer and a Temporal Aggregation Module (TAM) within the inference pipeline. In this setup, the query shot is sampled from one of the novel object classes present in the video domain. In our implementation, a target object from the first frame of the challenging VOT-LT video dataset—belonging to an unseen class—is designated as the one-shot query. Following this, the CD-BHRL-V network performs end-to-end cross-domain fine-tuning to improve adaptation to the temporal and appearance dynamics of the video domain. The trained model is then used to perform object detection on the remaining frames of the video, which collectively serve as the novel evaluation set.

In order to improve robustness to occlusion and appearance changes of the query object across the video frames, we augment the one-shot finetuning with a novel Online Target Update (OTU) mechanism controlled by a temporal aggregation module (TAM). To handle requirements of a long term online video object detection, the proposed CD-BHRL-OTU enables an unsupervised one-shot target update every K frames. Here unsupervised means the one-shot query update is performed by the detected video object bounding boxes through the online processing. K is a hyperparameter that needs to be specified depending on the application and the desired processing speed. When K is set to 1 CD-BHRL-OTU performs object detection after finetuning at each frame. However to minimize the processing load without reducing the object detection accuracy, we control the target update frames by TAM.

In this strategy, detected objects with confidence scores exceeding a predefined threshold thr_C are considered as one-shot query candidates. For each candidate, the Intersection over Union (IoU) with the latest query bounding box is monitored by TAM to leverage temporal information in video. The one-shot query update is triggered when IoU remains less than a threshold that shows the temporal consistency is disrupted. To guarantee an effective long term video object detection, every K frames, the OTU mechanism re-initializes the one-shot finetuning by updating the query with the candidate having the highest score (most confident detection) or if there is no candidate with the initial query sample.

IV. PERFORMANCE EVALUATION

Our code is implemented in Pytorch on top of the official code of the baseline architecture BHRL-C[1]. All evaluations are conducted with the GeForce RTX 4090 GPU. Excluding the finetuning, the accomplished inference rate on a 1280x720 video is 21.2 frames per second. Conventional AP50 and mAP metrics are used in the performance evaluation. For AP50, predicted bounding boxes that meet the IoU ≥ 0.5 condition are considered true positives (TP) and AP50 is calculated by finding the area under the precision-recall curve obtained for these bounding boxes. The average of all AP values for ten IoU thresholds from ≥ 0.50 to ≥ 0.95 is reported as mAP.

A. Comparison with SOTA on Still Images

As of our knowledge this work is the first attempt on the cross-domain OSVOD. Therefore we evaluated cross-domain performance on still image datasets to compare it with the state-of-the-art (SOTA) detectors. Following the setting used in [7], [8], CD-OSOD performance of BHRL-C (CD-BHRL-C) is reported on datasets ArTaxOr, UODD and DIOR, declared as the still image cross domain benchmarking datasets in [7], [8]. For this evaluation, we used the original class-based BHRL-C configuration as presented in [2], hence reported the mAP scores achieved over 5 runs with the base training COCO model. The mAP scores reported in Table I demonstrate that CD-BHRL-C outperforms or achieves comparable scores with the SOTA methods listed in the table, without finetuning. It is important to note that all of the listed SOTA methods except [7] and [12] are transformer-based architectures with a significantly higher number of learnable parameters compared to 48.4M parameters of CD-BHRL-C. This encouraged us to select BHRL-C as the baseline method in the design of our video object detector. In order to demonstrate the impact of finetuning we repeated the same test but performing a class-based finetuning on the query still images. The CD-BHRL-C-FT mAP scores reported at Table I highlight the performance improvement achieved by the proposed finetuning mechanism.

TABLE I: Cross-Domain OSOD mAP scores for the novel object classes where the base training is performed on COCO.

Methods	ArTaxOr	UODD	DIOR
Distill-cdfsod [7]	5.1	5.9	10.5
Detic [12]	0.6	0.0	0.1
Detic-FT [12]	3.2	4.2	4.1
DeViT [3]	0.4	1.5	2.7
DeViT-FT [3]	10.5	2.4	14.7
ViTDeT-FT [13]	5.9	4.0	12.9
CD-ViTO [8]	21.0	3.1	17.8
CD-BHRL-C(Ours)	**14.4**	**4.1**	**13.0**
CD-BHRL-C-FT (Ours)	**17.6**	**3.8**	**14.2**

B. Online Cross-domain Performance on Video

To evaluate the proposed CD-BHRL-OTU on the video, we have adopted to the online video object detection. The detector without target query update is referred to as CD-BHRL-FT. Both CD-BHRL-FT and CD-BHLR-OTU are finetuned with the query shot specified as the ground truth object in the first frame. For a fair comparison, following the setup in [2], the evaluations are performed using BHRL-C base training models

[1] https://github.com/hero-y/BHRL

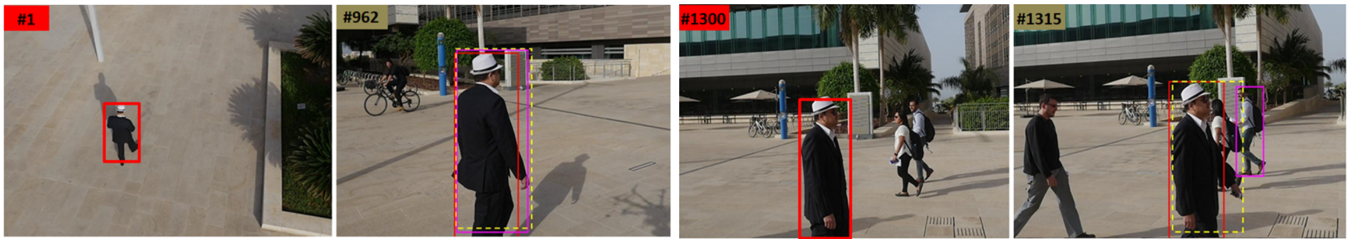

Fig. 2: Impact of the query update on Person20 video sequence where the person is a novel object class for the base training on COCO split-3 . CD-BHRL-FT(magenta), CD BHRL-OTU (yellow), GT (red). Unsupervised query update is executed at frame 1300.

TABLE II: Cross-domain AP50 and mAP scores reported on VOT-LT videos for the base training on different COCO splits.

	Method	Unseen (Novel)				
		split 1	split 2	split 3	split 4	Avg
AP50	CD-BHRL-V$_{1st}$	32.85	53.05	51.72	53.53	47.72
	CD-BHRL-FT	66.05	60.18	60.09	60.33	61.66
	CD-BHRL-OTU	69.80	66.23	69.81	63.65	67.37
mAP	CD-BHRL-V$_{1st}$	16.64	26.82	26.10	27.71	24.32
	CD-BHRL-FT	35.25	31.77	34.67	32.48	33.51
	CD-BHRL-OTU	36.03	37.66	37.01	31.88	35.65

TABLE III: Cross-domain AP50 and mAP scores reported on VOT-LT videos for the base training on VOC image dataset.

Method	Unseen (Novel)	
	AP50	mAP
CD-BHRL-V$_{1st}$	11.49	3.85
CD-BHRL-FT	58.05	29.82
CD-BHRL-OTU	64.20	31.03

trained on the four splits of the MS-COCO dataset, as well as the Pascal VOC dataset. In order to demonstrate object detection performance of the vanilla BHRL-C on video, the same query shot is fed into the network and detections across the video frames are evaluated. It is referred to as CD-BHRL-V$_{1st}$ to distinguish it from BHRL-C.

The VOT-LT benchmarking video dataset [11] is used for the evaluation in a cross-domain one-shot object detection setting. Each 50 videos of VOT-LT dataset has different numbers of frames, ranging from 1100 and 26277. When it is active, a 100 iterations finetuning is applied on the query shot. The target query update is triggered on the detections having at least 0.85 confidence score whenever the IoU between the best detection and the last target query bounding box remains less than 70%. Due to the diverse scene dynamics, the number of target updates varies significantly across videos. Note that in a video with L frames, the update mechanism can be activated at most L/K times where K is set to 100 for the reported results. The number of target updates is reported as 25.9 in average, with a value per video is ranging from 2 to 163.

Video object detection performance achieved with the base training on COCO image dataset is reported at Table II. AP50 and mAP scores achieved on the novel (unseen) classes of VOT-LT videos are reported. Results demonstrate that the integrated finetuning mechanism of CD-BHRL-FT provides 13.94% increase on average AP50 compared to CD-BHRL-V$_{1st}$. Impact of the proposed target update mechanism is an extra 5.71% gain achieved by CD-BHRL-OTU. Gain on average mAP achieved by CD-BHRL-FT compared to CD-BHRL-V$_{1st}$ is reported as 9.19%. CD-BHRL-OTU provides an extra 2.14% increase. We have also reported the detection performance on the seen object classes, means the target video object classes overlapping with the image classes. Respectively 3.91% and 2.68% higher average AP50 and mAP scores are achieved by CD-BHRL-OTU. This is because diversity of the video objects compared to the still images.

In order to evaluate the generalization capability of the proposed framework, we have tested the performance with a different baseline model, specifically the model generated by the base training on VOC image dataset. Table III demonstrates the detection performance increase achieved by the proposed detectors. Specifically, CD-BHRL-FT provides 46.56% increase on average AP50 compared to CD-BHRL-V$_{1st}$. Impact of the proposed target update mechanism is an extra 6.15% gain achieved by CD-BHRL-OTU. Gain on average mAP achieved by CD-BHRL-FT compared to CD-BHRL-V$_{1st}$ is reported as 25.97%. CD-BHRL-OTU provides an extra 1.21% increase. Similar to the base model trained on COCO image dataset, base training on VOC image dataset is successfully transferred to the video domain by the proposed CD-BHRL-FT and CD-BHRL-OTU architectures. Additionally, impact of the finetuning by CD-BHRL-FT is higher than the COCO model case. In our opinion, this is mainly because of the number of object classes in VOC is smaller than COCO that makes the finetuning with the query video shot crucial in domain adaptation.

Video frames illustrated in Figure 2 visually demonstrate impact of the Online Target Update mechanism. As it can be seen from frame 962 results, both CD-BHRL-FT and CD-BHRL-OTU well localize the target object bounding box specified in the first frame. However because of the excessive scale and pose changes across the video frames, CD-BHRL-FT fails to track the target object at frame 1315, while CD-BHRL-OTU robustly detects the target as a consequence of the target query update at frame 1300.

V. Conclusion

We propose a cross-domain one-shot inference architecture for the online video object detection applications. Differ from the existing work, we aim to transfer the base learning from image to video domain. The proposed fine-tuning scheme, supported by an unsupervised target shot update mechanism, can be seamlessly integrated into existing baseline one-shot detectors, facilitating adaptation to the video domain. The achieved AP50 and mAP scores demonstrate that the proposed framework has a potential to highly increase the novel video object detection rates without an additional base training thus alleviates a substantial reason of drastic performance drops in video object tracking-by-detection. Supplementary material is available at https://github.com/msprITU/CD-BHRL-OTU.

Acknowledgement: Authors thank to İrem B. Onur for her work at the initial step of this research.

References

[1] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E. Gonzalez, and Fisher Yu, "Frustratingly simple few-shot object detection," in *Proc. ICML*, 2020.

[2] Hanqing Yang, Sijia Cai, Hualian Sheng, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Yong Tang, and Yu Zhang, "Balanced and hierarchical relation learning for one-shot object detection," in *Proc. IEEE/CVF CVPR*, 2022, pp. 7581–7590.

[3] Xinyu Zhang, Yuhan Liu, Yuting Wang, and Abdeslam Boularias, "Detect everything with few examples," arXiv:2309.12969, 2023.

[4] Qi Fan, Chikeung Tang, and Wing Yu Tai, "Few-shot video object detection," in *Proc. ECCV*, 2022.

[5] Xudong Yao and Xiaoshan Yang, "Self-supervised spatial–temporal feature enhancement for one-shot video object detection," *Neurocomputing*, vol. 600, pp. 128219, 2024.

[6] Yogesh Kumar, Saswat Mallick, Anand Mishra, Sowmya Rasipuram, Anutosh Maitra, and Roshni Ramnani, "Qdetrv: Query-guided detr for one-shot object localization in videos," in *Proc. AAAI*, 2024, pp. 2831–2839.

[7] Wuti Xiong, "Cd-fsod: A benchmark for cross-domain few-shot object detection," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[8] Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qiu, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang, "Cross-domain few-shot object detection via enhanced open-set object detector," in *Proc. ECCV*, 2024, pp. 247–264.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft coco: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.

[10] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[11] Matej Kristan and et.al, "The ninth visual object tracking vot2021 challenge results," in *2021 IEEE/CVF ICCVW*, 2021, pp. 2711–2738.

[12] Zhou Xingyi, Girdhar Rohit, Joulin Armand, Kr¨ahenb¨uhl Philipp, and Ishan Misra, "Detecting twenty-thousand classes using image-level supervision," in *Proc. ECCV*, 2022, pp. 350–368, Springer.

[13] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He, "Exploring plain vision transformer backbones for object detection," in *Proc. ECCV*, 2022, pp. 280–296, Springer.