Supplementary Material for BoxMask: Revisiting Bounding Box Supervision for Video Object Detection

Khurram Azeem Hashmi Alain Pagani Didier Stricker Muhammamd Zeshan Afzal

DFKI - German Research Center for Artificial Intelligence, Kaiserslautern

firstname[0]_firstname[1].lastname@dfki.de

Overview. The supplementary material is organized as follows: Section 1 explains additional details of the employed network architecture; Section 2 presents ablation on the weightage of BoxMask loss. Section 3 provides more qualitative analysis of our method.

1. Network Architecture

1.1. Backbone Network

In the *conv5* block, we perform two modifications. First, we change the effective stride from 32 pxiels to 16 pxiels. Second, we replace the conventional 3×3 convolutions with dilated convolutions [3].

1.2. First Stage (Region Proposal Network)

The output of $conv4$ in the backbone network is propagated to the RPN to generate proposals. We use a total of 12 anchors consisting of 4 scales $64^2, 128^2, 256^2, 512^2$ and 3 aspect ratios $\{$1:2, 1:1, 2:1$\}$ in RPN. Adopting existing methods [2, 1], for each frame, we select 6000 proposals that pass the Non-Maximum Suppression (NMS) with an IoU threshold of 0.7 to obtain the top 300 proposals.

1.3. Second Stage and BoxMask Module

Instead of conventional RoIAlign, we apply temporal RoIAlign on the output of $conv5$ to extract aggregated RoI features from target and support frames. We set the identical hyperparameters in the temporal RoIAlign block and bounding box head as proposed in [1]. For BoxMask head, the spatial resolution $(m \times m)$ for each sampled RoI is set to 14×14. Later, these temporal RoI features are propagated to BoxMask head to perform mask prediction in parallel to classification and regression.

2. More Ablation Studies

2.1. Hyperparameters for BoxMask

We investigate different weights λ for our BoxMask loss L_{bm} by incorporating the BoxMask module in TROI [1].

Weight (λ)	0.0 (TROI)	0.1	0.3	0.5	0.7	0.9	1.0
mAP (%)	78.9	79.0	80.1	**80.7**	80.4	80.1	79.9

Table 1. Effect of increasing weight for BoxMask loss when BoxMask is equipped in TROI on ResNet-50. Best result is highlighted in bold.

As summarized in Table 1, the mAP tends to increase with the rise in λ and achieves best results of 80.7 when (λ) is set to 0.5. After appreciating the weight further, we observe a decline in mAP. We argue that considering the aim of our BoxMask head is to assist classification and localization, assigning higher weight ($\lambda > 0.5$) for L_{bm} leads to overfitting. Furthermore, since the ground truth BoxMask tensor, M_{box} contains both object and background, increasing (λ) forces the network to learn high-level features, resulting in misclassification. Therefore, we set $\lambda = 0.5$ in all our experiments.

3. Additional Visualizations

3.1. t-SNE Visualization for all Categories on ImageNet VID

In order to demonstrate how the proposed BoxMask module alleviates object confusion in SELSA [2], we extract the learned proposal features prior to detection and visualize them with t-SNE in Figure 1. The left part of the figure shows clusters of learned proposal features with instance-level feature aggregation, while the right part depicts clusters of both class-aware pixel-level learning and instance-level feature aggregation. As shown in Figure 1, by employing instance-level feature aggregation only in SELSA, proposals from objects who either look-alike such as *domestic_cat* and *dog* or objects with similar motion characteristics such as *car*, *bus*, and *watercraft* are misplaced into an incorrect cluster. On the contrary, when class-aware pixel-level learning is incorporated into instance-level feature aggregation, we observe compact clusters with severely reduced misclassification of object

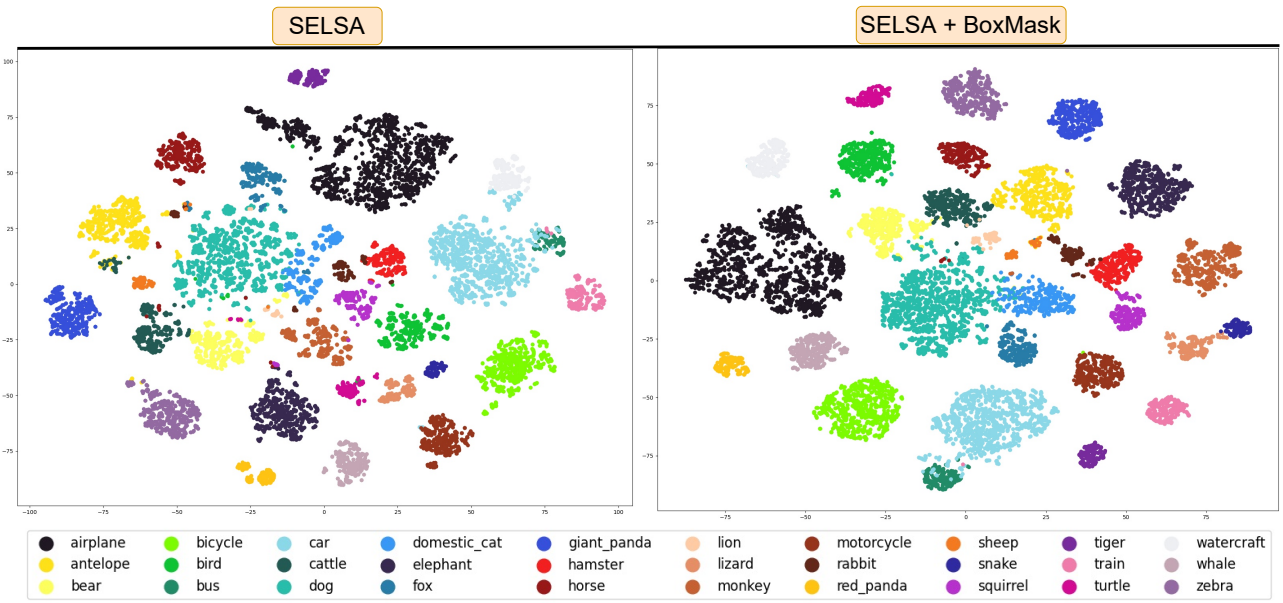

Figure 1. t-SNE visualization of learned proposal features with and without our BoxMask module in SELSA [2]. With instance-level feature aggregation only in SELSA, objects sharing similar motion attributes (*bus*, *car*, and *watercraft*) or with similar appearances (*dog*, and *domestic_cat*) are confused with each other. The class-aware pixel-level learning in our BoxMask module introduces discriminative cues which alleviate the object confusion and enable the proposal features to form compact clusters, as shown in SELSA+BoxMask. Best view in color.

Figure 2. Qualitative analysis between existing methods without and with our BoxMask module on the validation set of the ImageNet VID dataset under different scenarios. With our BoxMask in SELSA [2], the detector removes false classification (turtle) and provides consistent results in case of rare poses and part-occlusions (top four rows). With the addition of our BoxMask module in TROI [1], the combination of temporal and semantic cues yields stable detections that are more robust to the motion blur/camera defocus and part-occlusions (last four rows). The illustration scheme is inspired by [4]. Best view it on the screen and zoom in.

proposals.

3.2. More Visual Detection Results

Figure 2 depicts an extended version of the qualitative analysis presented in the main paper. When our BoxMask module is plugged into recent state-of-the-art methods [2, 1], consistent and significant improvements are observed even under the challenging scenarios of video object detection.

References

[1] Tao Gong, Kai Chen, Xinjiang Wang, Qi Chu, Feng Zhu, Dahua Lin, Nenghai Yu, and Huamin Feng. Temporal roi align for video object recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1442–1450, 2021.

[2] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9217–9225, 2019.

[3] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[4] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: End-to-end video object detection with spatial-temporal transformers. *arXiv preprint arXiv:2201.05047*, 2022.