Future Object Detection Using Frame Prediction

Taiju Watanabe
Graduate School of FSE, Waseda University
Tokyo, Japan

Takahiro Shindo
Graduate School of FSE, Waseda University
Tokyo, Japan

Kein Yamada
School of FSE, Waseda University
Tokyo, Japan

Hiroshi Watanabe
Graduate School of FSE, Waseda University
Tokyo, Japan

***Abstract*—Future object detection is a task in computer vision that predicts bounding box coordinates, confidence score and class prediction of the objects in the future frames from the past few frames of a video. In this task, not only detecting objects, but also acquiring temporal information is required. To tackle this problem, we propose one-stage detection model based on YOLOV. We improve the detection performance of YOLOV by incorporating the image restoration technique, called Dual Attention Block. Experimental results show that our model achieves better performance than YOLOV (up to 6% gain in mAP) on several settings of future object detection.**

***Keywords*—YOLOV, Dual Attention Block**

I. Introduction

To understand the valuable information of the image, understanding the content and location of the image is necessary. This task is referred to as object detection.

Video object detection is a task in computer vision that predicts the location and class of objects in the video. Due to the fact that objects in videos often contain motion blur, sudden occlusion and camera defocus, object detection in a video is considered more difficult than object detection in an image. To address these difficulties, several models have been proposed, such as YOLOV [1].

Future object detection is a task that predicts bounding box coordinates, confidence and class prediction of the object of future frames from the past few frames of a video. Future object detection can be applied to risk prediction in autonmous driving. Future object detection is difficult than video object detection since acquiring temporal information is required. Future object detection can be interpreted as follows. Given frames from $t = 1$ to $t = T$ ($\{\mathbf{I}^T\}_{t=1}^{T}$), detection model $\mathbf{F}$ predicts the vector $\mathbf{O}^{T+\tau}$ containing bounding box coordinates, confidence and class prediction of objects in the frames $\{\mathbf{I}^T\}_{t=T+1}^{\tau}$. This is represented as

$$\mathbf{O}_{T+1}^{T+\tau} = \mathbf{F}(I^1, I^2, I^3, \cdots, I^T). \tag{1}$$

In this paper, we use YOLOV as a base detector to tackle future object detection. We incorporate image restoration method, called Dual Attention Block [2] to acquire temporal information and to predict future frames. We further improve the model by incorporating feature loss of YOLOV.

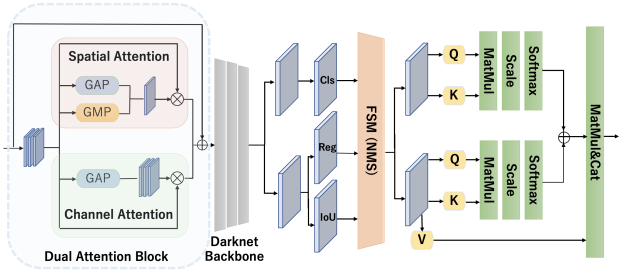

Fig. 1. Model structure of proposed method

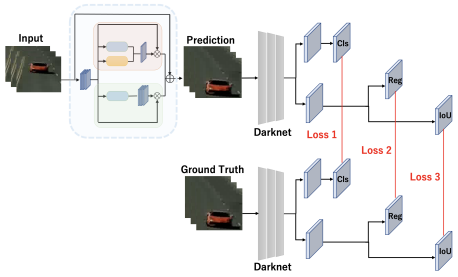

Fig. 2. Darknet Feature Loss

II. Related Work

YOLOV is a video detection model which uses YOLOX [3] as a base detector. YOLOX extracts features from the input frames. According to the prediction of YOLOX, the Feature Selection Module (FSM) selects top k features with high confidence scores. With these features, non-maximum suppression (NMS) is applied for further refinement. Finally, Feature Aggregation Module (FAM) is applied for the classification. YOLOV marks the state-of-the art performance on video object detection tasks among one-stage detectors.

III. Proposed Method

A. Network Architecture

YOLOV is developed for video object detection, and it does not acquire temporal information. In order to obtain temporal information, we predict future frames using video refinement module. Our model architecture is shown in Fig. 1. As for the refinement module, we use Dual Attention Block

979-8-3503-4017-4/23/$31.00 ©2023 IEEE

(DAB). DAB was first presented in CycleISP, well known denoising model. DAB uses two attention mechanisms, spatial attention and channel attention. Spatial attention is used to acquire spatial relationships of features and computes a spatial attention map. Spatial attention map is obtained using the combination of global average pooling (GAP) and global max pooling (GMP). Channel attention is used to exploit channel relations of convolutional features. It uses GAP to extract features of the channel. By using DAB, it is possible to predict future frames.

B. Darknet Feature Loss (DF Loss)

To obtain features of future frames, we propose Darknet Feature Loss. Detail is described in Fig. 2. By applying predicted frames and ground truth frames to Darknet backbone (D), we can obtain features corresponding to class probability, bounding box prediction and IoU. We use the mean squared difference between features of predicted frames and ground truth frames for our loss function. It is described as follows. Predicted frames refer to $\mathbf{I}_P^T$, ground truth frames refer to $\mathbf{I}_{GT}^T$.

$$DFLoss = MSE(D(\mathbf{I}_P^T), D(\mathbf{I}_{GT}^T)). \quad (2)$$

This loss is added to the regular loss of YOLOV.

IV. Experiment

We use the videos in the ImageNet VID [4] and the ImageNet DET with the same classes as our training and evaluation data. The ImageNet VID dataset is a dataset for video object detection. It contains 3,862 videos for training and 555 videos for evaluation. There are 30 categories in the ImageNet VID dataset. We compared our model and YOLOV on three tasks of future object detection, predicting objects in 2 frames from the past 2 frames, detection in 3 frames from the past 3, detection in 4 frames from the past 4. We initialized our model and YOLOV from the ImageNet VID pretrained weights. Dual Attention Block, the linear projection layers in YOLOX and the attention mechanism in YOLOV is finetuned for our model. As for comparison, linear projection layers and the attention mechanism is finetuned for YOLOV. Our model and YOLOV is finetuned for 10 epochs. When training, videos are randomly resized from 352x352 to 672x672. For testing, videos are uniformly resized to 576x576. As for the evaluation metrics, we use AP, AP50, AP75 and AR.

V. Result

Experimental results are shown in Table I-III. There are three models for YOLOV and our proposed method depending on the number of parameters. S, L, X corresponds to the number of parameters in the increasing order. YOLOV-(S,L,X)+DDB corresponds to our proposed method. These models are equipped with Dual Attention Block and is trained with loss functions which Darknet Feature Loss is combined. From the result, our model shows better performance for predicting objects in future 2 or 3 frames (up to 6% gain in mAP). Also, our model shows competitive performance for predicting objects in future 4 frames. Therefore, frame prediction is effective for future object detection.

TABLE I
Results of future object detection (2 to 2). The optical results are marked by **bold**.

Model	AP	AP50	AP75	AR
YOLOV-S	42.5	67.2	44.9	56.7
YOLOV-S+DAB	**43.2**	**68.3**	**45.6**	**57.6**
YOLOV-L	50.6	74.6	55.6	**63.4**
YOLOV-L+DAB	**51.0**	**74.9**	**56.0**	63.2
YOLOV-X	52.4	76.8	57.2	63.7
YOLOV-X+DAB	**53.0**	**77.6**	**57.7**	**64.1**

TABLE II
Results of future object detection (3 to 3). The optical results are marked by **bold**.

Model	AP	AP50	AP75	AR
YOLOV-S	41.8	67.9	44.2	55.9
YOLOV-S+DAB	**44.4**	**69.3**	**48.3**	**58.4**
YOLOV-L	49.1	74.5	53.1	61.7
YOLOV-L+DAB	**52.5**	**76.6**	**58.2**	**64.7**
YOLOV-X	49.0	73.8	52.7	61.8
YOLOV-X+DAB	**53.2**	**76.9**	**58.9**	**65.6**

TABLE III
Results of future object detection (4 to 4). The optical results are marked by **bold**.

Model	AP	AP50	AP75	AR
YOLOV-S	**37.3**	**64.8**	**37.1**	**52.2**
YOLOV-S+DAB	37.0	**64.8**	36.4	51.6
YOLOV-L	45.0	**72.7**	47.1	**57.8**
YOLOV-L+DAB	**45.2**	72.5	**47.7**	57.4
YOLOV-X	45.3	72.3	47.9	57.7
YOLOV-X+DAB	**48.9**	**73.7**	**52.9**	**61.7**

VI. Conclusion

In this paper, we propose a method for improving the accuracy of future object detection using YOLOV. By predicting future frames using Dual Attention Block, we can obtain temporal information which was not capable for YOLOV. Moreover, our model is able to learn features of future frames by incorporating Darknet Feature Loss. Experimental results show that our model exceeds YOLOV in several future object detection tasks.

Acknowledgment

These research results were obtained from the commissioned research (No. 05101) by National Institute of Information and Communications Technology (NICT), Japan.

References

[1] Y. Shi, N. Wang and X. Guo, "YOLOV: Making Still Image Object Detectors Great at Video Object Detection," arXiv:2208.09686, 1-9, Mar. 2023.

[2] S. W. Zamir, A. Arora, S. Khan, M. Hayat and F. S. Khan, M. Yang, L. Shao, "CycleISP: Real Image Restoration via Improved Data Synthesis," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2693-2702, Jun. 2020.

[3] Z. Ge, S. Liu, F. Wang, Z. Li and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," arXiv:2107.0843, 1-7, Aug. 2021.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision (IJCV), vol. 115, 211-252, Apr. 2015.