

# YOWOv2: A Stronger yet Efficient Multi-level Detection Framework for Real-time Spatio-temporal Action Detection

Jianhua Yang<sup>1,2</sup>, Kun Dai<sup>1</sup>

**Abstract**—Designing a real-time framework for the spatio-temporal action detection task is still a challenge. In this paper, we propose a novel real-time action detection framework, YOWOv2. In this new framework, YOWOv2 takes advantage of both the 3D backbone and 2D backbone for accurate action detection. A multi-level detection pipeline is designed to detect action instances of different scales. To achieve this goal, we carefully build a simple and efficient 2D backbone with a feature pyramid network to extract different levels of classification features and regression features. For the 3D backbone, we adopt the existing efficient 3D CNN to save development time. By combining 3D backbones and 2D backbones of different sizes, we design a YOWOv2 family including YOWOv2-Tiny, YOWOv2-Medium, and YOWOv2-Large. We also introduce the popular dynamic label assignment strategy and anchor-free mechanism to make the YOWOv2 consistent with the advanced model architecture design. With our improvement, YOWOv2 is significantly superior to YOWO, and can still keep real-time detection. Without any bells and whistles, YOWOv2 achieves 87.0% frame mAP and 52.8% video mAP with over 20 FPS on the UCF101-24. On the AVA, YOWOv2 achieves 21.7% frame mAP with over 20 FPS. Our code is available on <https://github.com/yjh0410/YOWOv2>.

**Index Terms**—Spatio-temporal action detection, one-stage detection, spatial encoder, temporal encoder

## I. INTRODUCTION

**S**patio-temporal action detection (STAD) aims to detect action instances in the current frame. It has been widely applied, such as video surveillance [1] and somatosensory game [2].

Some researchers [3]–[5] employ 3D CNNs [6], [7] to extract spatio-temporal information from video clips to accurately detect actions, as the occurrence of actions is a continuous concept over time. However, many 3D CNN-based frameworks suffer from poor detection speeds, which prevent them from operating in real-time due to the massive computational requirements of the hefty 3D CNNs they use.

Hence, other researchers [8]–[10] leverage 2D CNNs [11], [12] to develop more efficient action detection frameworks. The key concept behind these 2D CNN-based frameworks is to use a parameter-sharing 2D CNN to extract spatial features frame-by-frame and store them in a buffer. Subsequently, they only need to process the new input frame, combine its spatial

features with those in the buffer, and generate the spatio-temporal features for the final detection. Nonetheless, such a pipeline cannot fully model temporal association, and real-time detection is only feasible with RGB streams. When optical flow is used, although performance improves, the speed is significantly reduced.

On the contrary, Köpüklü et al. [13] develops a novel one-stage action detector, You Only Watch Once (YOWO), by combining a 2D backbone [14] for spatial localization and a 3D backbone for spatio-temporal modeling. To mitigate the high computational cost of 3D CNNs, they designs a series of efficient 3D CNNs [15] as the 3D backbone for efficient inference. After the backbones, YOWO employs a channel encoder to fuse the two features for the final detection. With their designs, YOWO achieves excellent performance on popular benchmarks and is touted as a fast action detector. However, YOWO still suffers from two disadvantages:

- YOWO is a one-level detector and performs the final detection on a low-level feature map, impairing the detection performance for small action instances.
- YOWO is an anchor-based method and has lots of anchor boxes with many hyperparameters, such as the number, size, and aspect ratio of anchor boxes. Those hyperparameters must be carefully artificially designed, impairing the generalization.

In summary, **designing a real-time detection framework for spatio-temporal action detection remains a challenge.**

In this study, we propose YOWOv2, a brand-new real-time action detector. A 3D backbone with a multi-level 2D backbone make up YOWOv2. A multi-level detection pipeline is designed for YOWOv2 to detect action occurrences of various scales thanks to our multi-level 2D backbone with a feature pyramid network (FPN) [16]. We also recommend the quick deployment 3D CNNs [15] for the 3D backbone. The disadvantages of the anchor box are also avoided by using the anchor-free mechanism. We use a dynamic label assignment technique because the anchor box is removed, enhancing the adaptability of the YOWOv2. Moreover, we construct a variety of YOWOv2 models, such as **YOWOv2-Tiny**, **YOWOv2-Medium** and **YOWOv2-Large** by merging 3D backbones with 2D backbones of various sizes for platforms with diverse computing power.

Compared to YOWO, YOWOv2 delivers superior performance on the UCF101-24 [17] and AVA [18] datasets and boasts significant advantages in terms of both parameter count and FLOPs. Moreover, YOWOv2 is capable of real-time

This work was supported in part by the National Natural Science Foundation of China (62176072).

<sup>1</sup>Jianhua Yang and Kun Dai are with the State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin 150001, China.

<sup>2</sup>Jianhua Yang is with Wuhu Robot Industry Technology Research Institute, Harbin Institute of Technology, Wuhu 241000, China.

arXiv:2302.06848v2 [cs.CV] 8 Jun 2023

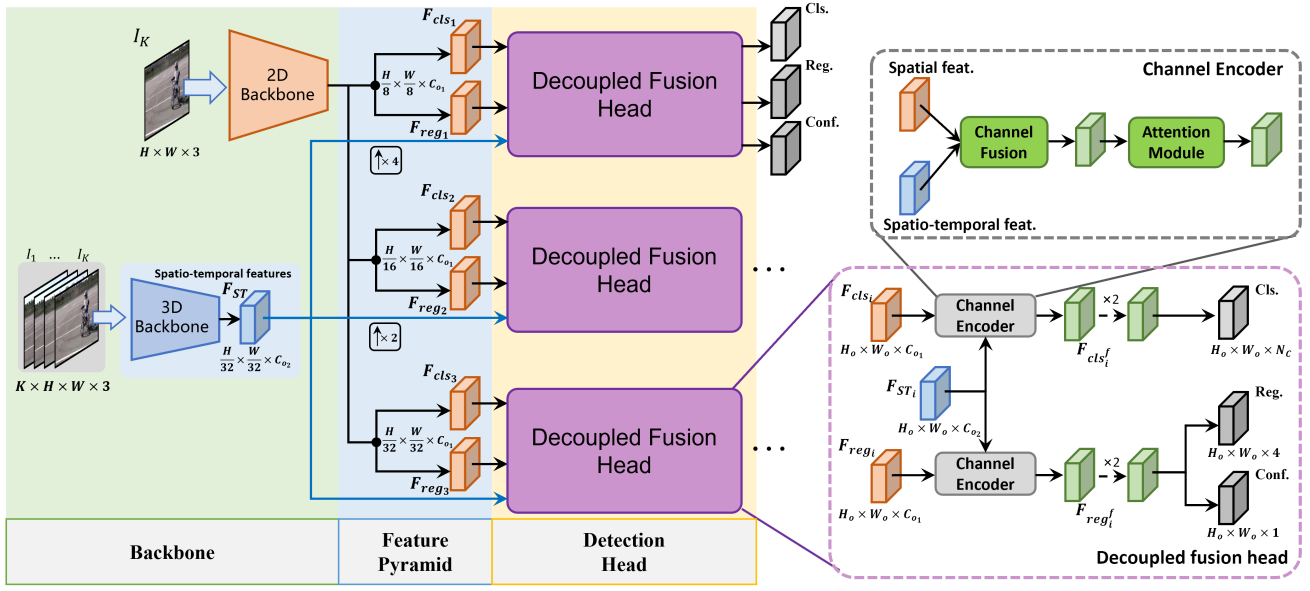


Fig. 1. Overview of YOWOv2. YOWOv2 uses upsampling operation to align the spatio-temporal features output by the 3D backbone with the spatial features of each level output by the 2D backbone and uses the Decoupled fusion head to achieve the fusion of the two features on each level. Finally, YOWOv2 outputs the multi-level confidence predictions, classification predictions, and regression predictions respectively.

operation. In comparison to other real-time action detectors, YOWOv2 also achieves better performance. In summary, our contributions are as follows:

- We propose a new real-time action detection framework, YOWOv2 with a multi-level detection structure, which is friendly to detect small action instances.
- YOWOv2 features an anchor-free detection pipeline, which eliminates the limitations of anchor boxes.
- We design a YOWOv2 family by combining the 3D backbones and 2D backbones of different sizes for the platforms with different computing power.
- YOWOv2 achieves state-of-the-art performance on popular benchmarks, compared to other real-time action detectors.

## II. RELATED WORK

### A. Spatio-temporal action detection

Spatio-temporal action detection involves detecting and identifying all instances of action that occur within a given frame. To achieve accurate action detection, it is essential to effectively extract spatio-temporal features.

**3D CNN-based methods.** Some researchers use the 3D CNN to design action detectors [3], [4], [19]–[22], due to the strong spatio-temporal modeling capabilities. Girdhar et al. [3] use the I3D [6] to generate action region proposals and then use the Transformer [23] to complete the final detection. Zhao et al. [5] deploy a 3D CNN to encode input video and then use the Transformer with the tuber queries for final detection. Although these 3D CNN-based methods achieve impressive success, they all suffer from the expensive computation of the heavy 3D CNN and are therefore too slow to run in real time.

**2D CNN-based methods.** Another approach is to separate spatio-temporal associations and design 2D CNN-based action

detectors for efficient detection. For instance, Kalogeiton et al. [8] devise a one-stage detection framework called ActionTubelet (ACT). They utilize SSD [11] to extract spatial features from each frame in a video clip and then merge them. Subsequently, a detection head is employed to process the merged spatial features for the final detection. Li et al. [10] follow the ACT framework and develop an anchor-free one-stage action detector called MovingCenter (MOC). Ma et al. [24] further enhance the MOC with a self-attention mechanism. However, the real-time detection performance of these methods can only be ensured when RGB streams are applied as input. When optical flow is added, their speed significantly declines, despite the improved performance. Moreover, obtaining high-quality optical flow requires offline processing, which cannot meet the demands of online operations.

## III. METHODOLOGY

### A. Preliminary

The overview of YOWOv2 is shown in Fig. 1. Given a video clip with  $K$  frames  $V = \{I_1, I_2, \dots, I_K\}$  where the  $I_K$  is the current frame, YOWOv2 uses an efficient 3D CNN [15] as the 3D backbone to extract spatio-temporal features  $F_{ST} \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times \frac{C_{o2}}{32}}$ . The 2D backbone of YOWOv2 is a multi-level 2D CNN, responsible for outputting decoupled multi-level spatial features  $F_{cls} = \{F_{cls_i}\}_{i=1}^3$  and  $F_{reg} = \{F_{reg_i}\}_{i=1}^3$  of  $I_K$ , where the  $F_{cls_i} \in \mathbb{R}^{\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times C_{o1}}$  is the classification features and  $F_{reg_i} \in \mathbb{R}^{\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times C_{o1}}$  is the regression features. After the two backbones, we deploy two channel encoders on each feature map of level to integrate features. After that, two extra parallel branches with two  $3 \times 3$  conv layers followed the channel encoders to predict  $Y_{cls_i} \in \mathbb{R}^{\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times N_C}$  for classification,  $Y_{reg_i} \in \mathbb{R}^{\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times 4}$  for regression respectively. A confidence branch is added

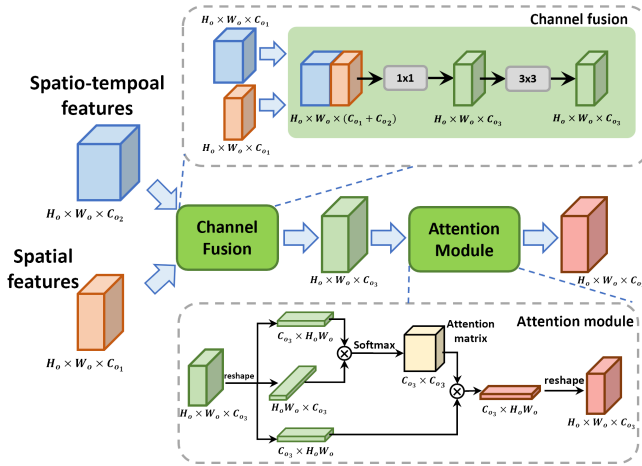


Fig. 2. Overview of ChannelEncoder. It contains the channel fusion and channel self-attention mechanism, which are both used to fuse 2D and 3D features.

on the regression branch to predict  $Y_{conf_i} \in \mathbb{R}^{\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times 1}$  for actionness confidence. Next, we introduce the design of YOWOV2 in detail.

### B. Design of YOWOV2

**2D backbone.** The 2D backbone is supposed to extract multi-level spatial features from the current frame. Considering the balance between performance and speed, we draw some advanced ideas from the advanced object detectors [25], [26]. We reuse the backbone and feature pyramid network (FPN) of YOLOv7 [25] to save training time. After the FPN, we add extra  $1 \times 1$  conv layers to compress the channel number of each level feature map  $F_{S_i}$  to  $C_{o1}$  which is defaulted to 256. Then, we add two parallel branches with two  $3 \times 3$  conv layers to output decoupled features, as shown in Eq.(1).

$$\begin{aligned} F_{cls_i} &= f_{conv2}^1 (f_{conv1}^1 (F_{S_i})) \\ F_{reg_i} &= f_{conv2}^2 (f_{conv1}^2 (F_{S_i})) \end{aligned} \quad (1)$$

where the  $f_{conv_j}^i$  is the  $j^{th}$   $3 \times 3$  conv layer of the  $i^{th}$  branch.

In YOWOV2 framework, the 2D backbone outputs the decoupled feature maps of three levels,  $F_{cls} = \{F_{cls_i}\}_{i=1}^3$  and  $F_{reg} = \{F_{reg_i}\}_{i=1}^3$ . We name the 2D backbone FreeYOLO for convenience. By controlling the depth and width of FreeYOLO, we designed two FreeYOLO of different sizes, FreeYOLO-Tiny for YOWOV2-Tiny and FreeYOLO-Large for YOWOV2-Medium and YOWOV2-Large. To accelerate the convergence of training, we pretrain our 2D backbone with additional  $1 \times 1$  conv layers on the COCO [27]. The pretrained weight files are available on the GitHub<sup>1</sup>.

**3D backbone.** The 3D backbone is supposed to extract the spatio-temporal features  $F_{ST}$  from the video clip for the spatio-temporal association. We deploy the efficient 3D CNN [15] to reduce computation and thus guarantee real-time

detection. To fuse with decoupled spatial features, we simply upsample  $F_{ST}$  to obtain  $\{F_{ST_i}\}_{i=1}^3$ , as shown in Eq.(2).

$$\begin{aligned} F_{ST_1} &= \text{Upsample}_{4 \times} (F_{ST}) \\ F_{ST_2} &= \text{Upsample}_{2 \times} (F_{ST}) \\ F_{ST_3} &= F_{ST} \end{aligned} \quad (2)$$

where the Upsample is the upsampling operation for aligning  $F_{ST_i} \in \mathbb{R}^{\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times C_{o2}}$  and  $F_{cls_i}$  and  $F_{reg_i}$  in the spatial dimension.

**ChannelEncoder.** ChannelEncoder, proposed by YOWO [13], is supposed to fuse the features from the 2D backbone and 3D backbone. Given a  $F_S \in \mathbb{R}^{H_o \times W_o \times C_{o1}}$  and a  $F_{ST} \in \mathbb{R}^{H_o \times W_o \times C_{o2}}$ , the ChannelEncoder first concatenates them along the channel dimension and uses two naive conv layer followed a BN and LeakyReLU to achieve primary channel integration, as following,

$$F_f = f_{conv2} (f_{conv1} (\text{Concat} [F_S, F_{ST}])) \quad (3)$$

where the  $F_f \in \mathbb{R}^{H_o \times W_o \times C_{o3}}$ , Concat is the channel concatenation operation,  $f_{conv1}$  and  $f_{conv2}$  are both the conv layers with BN and LeakyReLU. Then, the  $F_f$  is reshaped to  $F_{f_2} \in \mathbb{R}^{C_{o3} \times H_o W_o}$  for the following channel self-attention mechanism inspired by DANet [28] to do deeper processing, so that the information containing two different levels features can be fully integrated, as shown in Eq.(4)

$$F_{f_3} = \text{Softmax} (F_{f_2} F_{f_2}^T) F_{f_2} \quad (4)$$

Finally, the  $F_{f_3} \in \mathbb{R}^{C_{o3} \times H_o W_o}$  is reshaped to  $F_f \in \mathbb{R}^{H_o \times W_o \times C_{o3}}$  followed by another conv layer. The whole pipeline of the ChannelEncoder is shown in Fig.2.

**Decoupled fusion head.** In YOWOV2, the 2D backbone outputs the decoupled spatial features  $F_{cls} = \{F_{cls_i}\}_{i=1}^3$  and  $F_{reg} = \{F_{reg_i}\}_{i=1}^3$  of the current frame  $I_K$  while the 3D backbone outputs  $\{F_{ST_i}\}_{i=1}^3$  obtained by upsampling  $F_{ST}$  of the video clip  $V = \{I_1, I_2, \dots, I_K\}$ . Note that  $F_{cls_i}$  and  $F_{reg_i}$  contain very different semantic information, which inspires us to fuse  $F_{cls_i}$  and  $F_{reg_i}$  with  $F_{ST_i}$  separately. Therefore, we design a decoupled fusion head to fuse  $F_{ST_i}$  into  $F_{cls_i}$  and  $F_{reg_i}$  independently, as shown in Eq.(5).

$$\begin{aligned} F_{cls_i}^f &= \text{ChannelEncoder} (F_{cls_i}, F_{ST_i}) \\ F_{reg_i}^f &= \text{ChannelEncoder} (F_{reg_i}, F_{ST_i}) \end{aligned} \quad (5)$$

After the feature aggregation, we deploy two parallel branches on each level for final detection. Its design is simple, just a classification branch and a box regression branch.

For the classification branch, it outputs the classification prediction  $Y_{cls_i} \in \mathbb{R}^{\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times N_C}$ , where  $Y_{cls_i}(x, y)$  represents the probability of action instances at each spatial position on  $Y_{cls_i}$  and  $N_C$  is the number of action classes. Taking  $F_{cls_i}^f$ , the branch applies two  $3 \times 3$  conv layers, each with  $C$  filters and each followed by SiLU activations. Finally, a  $1 \times 1$  conv layer with  $N_C$  filters and sigmoid activations is attached to output the  $N_C$  binary predictions per spatial position.

For the box regression branch, it outputs the box regression prediction  $Y_{reg_i} \in \mathbb{R}^{\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times 4}$ , where  $Y_{reg_i}(x, y)$  represents the 4 relative offsets at each spatial position. The design

<sup>1</sup><https://github.com/yjh0410/FreeYOLO>

is equal to the classification branch except that the final  $1 \times 1$  conv layer is with 4 filters for offset predictions. Additionally, an extra  $1 \times 1$  conv layer with 1 filter is added into this branch for actionness confidence prediction,  $Y_{conf_i} \in \mathbb{R}^{\frac{H}{2^{t+2}} \times \frac{W}{2^{t+2}} \times 1}$ . Note that, there is no anchor box in each spatial position, therefore, YOWOv2 is an anchor-free method.

### C. Label assignment

Since YOWOv2 is an anchor-free action detector without any anchor boxes, the multi-level label assignment becomes a challenge. Recently, dynamic label assignment has shown success in object detection. Inspired by YOLOX [26], we implement SimOTA for the label assignment of YOWOv2. Specifically, we calculate the cost between all predicted bounding boxes and groundtruths. Eq.(6) demonstrates the cost between the  $i^{th}$  prediction and the  $j^{th}$  ground truth. Subsequently, each groundtruth is assigned with the  $top_k$  predicted bounding boxes with the least cost, where  $k$  is dynamically determined by the IoU between the predicted bounding boxes and the target bounding boxes.

$$c_{ij}(\hat{a}_i, a_j, \hat{b}_i, b_j) = L_{cls}(\hat{a}_i, a_j) + \gamma L_{seg}(\hat{b}_i, b_j) \quad (6)$$

where the  $\hat{a}_i$  and  $a_j$  are the classification prediction (multiplied by confidence prediction) and target,  $\hat{b}_i$  and  $b_j$  are the regression prediction and target and  $\gamma$  is the cost balance factor, empirically being 3 in the experiments.

### D. Loss function

We define loss function as follows:

$$\begin{aligned} L(\{a_{x,y}\}, \{b_{x,y}\}, \{c_{x,y}\}) &= \frac{1}{N_{pos}} \sum_{x,y} L_{conf}(\hat{c}_{x,y}, c_{x,y}) \\ &+ \frac{1}{N_{pos}} \sum_{x,y} \mathbb{I}_{\{\hat{a}_{x,y} > 0\}} L_{cls}(\hat{a}_{x,y}, a_{x,y}) \\ &+ \frac{\lambda}{N_{pos}} \sum_{x,y} \mathbb{I}_{\{\hat{a}_{x,y} > 0\}} L_{reg}(\hat{b}_{x,y}, b_{x,y}) \end{aligned} \quad (7)$$

where  $L_{conf}$  and  $L_{cls}$  are both the binary cross-entropy and  $L_{reg}$  is the GIoU loss [29]. The  $a_{x,y}$ ,  $b_{x,y}$  and  $c_{x,y}$  are classification prediction, regression prediction, and confidence prediction, while the  $\hat{a}_{x,y}$ ,  $\hat{b}_{x,y}$  and  $\hat{c}_{x,y}$  are groundtruths.  $N_{pos}$  denotes the number of positive samples and  $\lambda$  is the loss balance factor, empirically being 5 in the experiments.  $\mathbb{I}_{\{\hat{a}_{x,y} > 0\}}$  is the indicator function, being 1 if  $\hat{a}_{x,y} > 0$  and 0 otherwise.

## IV. EXPERIMENTS

### A. Datasets

**UCF101-24** [17]. UCF101-24 contains 3,207 untrimmed videos for 24 sports classes and provides corresponding spatio-temporal annotations. There may be multiple action instances per frame. Following YOWO [13], we train and evaluate YOWO-Plus on the first split.

**AVA** [18]. AVA is a large-scale benchmark for spatial-temporal action detection. It contains 430 15-minute video

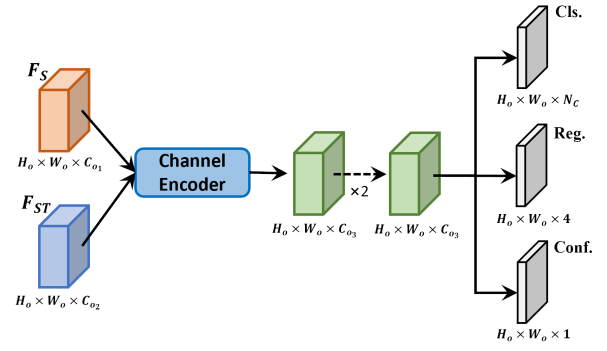


Fig. 3. Coupled fusion head. In the coupled head, the spatial features from the 2D backbone is also coupled which means that the parallel  $3 \times 3$  conv layers after the FPN are removed.

TABLE I  
PERFORMANCE COMPARISON BETWEEN COUPLED FUSION HEAD (CFH) AND DECOUPLED FUSION HEAD (DFH) ON THE UCF101-24.

Head	Model	FPS	F-mAP (%)	V-mAP (%)
CFH	YOWOv2-T	56	78.9	49.8
	YOWOv2-M	45	81.2	50.7
	YOWOv2-L	33	84.3	51.5
DFH	YOWOv2-T	50	80.5	51.3
	YOWOv2-M	42	83.1	50.7
	YOWOv2-L	30	<b>85.2</b>	<b>52.0</b>

clips with 80 atomic visual actions (AVA). It provides annotations at 1 Hz in space and time, and precise spatio-temporal annotations with possibly multiple annotations for each person. Therefore, this benchmark is very challenging. We train YOWOv2 on the train split and evaluate it on the most-frequent 60 action classes of the AVA dataset. We report evaluation results on the AVA v2.2.

### B. Implementation details

For training, we use the AdamW optimizer with an initial learning rate 0.0001 and weight decay 0.0005. The batch size is set to 8 with 16 gradient accumulate. On the UCF101-24, we train YOWOv2 for 7 epochs and decay the learning rate by a factor of 2 at 1, 2, 3, and 4 epoch, respectively. On the AVA, we train YOWOv2 for 9 epochs and decay the learning rate by a factor of 2 at 3, 4, 5, and 6 epoch, respectively. Unless otherwise specified, the size of the input frame is reshaped to  $224 \times 224$ .

For evaluation metrics, we follow previous works [5], [13], [21] to report frame mAP (F-mAP) and video mAP (V-mAP) at 0.5 IoU between predictions and groundtruths. We follow the link algorithm of YOWO [13] to build action tubelets. On the AVA, we report frame mAP at 0.5 IoU since the annotations are sparsely provided at 1 Hz.

### C. Effectiveness of decoupled fusion head

To evaluate the impact of the decoupled fusion head on YOWOv2, we design a coupled fusion head as a control group, as shown in Fig.3. We conduct experiments on UCF101-24

TABLE II  
ABLATION STUDY OF THE EFFECTIVENESS OF LOSS BALANCE FACTOR  $\lambda$ .

$\lambda$	UCF101-24		AVA
	F-mAP (%)	V-mAP (%)	mAP (%)
1.0	83.3	50.1	19.6
2.0	84.7	51.1	19.8
3.0	85.1	51.9	20.0
4.0	85.2	52.0	20.2
5.0	85.2	52.0	20.2
6.0	85.0	52.0	20.1
7.0	84.8	51.8	20.0

TABLE III  
COMPARISON WITH YOWO ON THE UCF101-24. FPS IS MEASURED ON A GPU RTX 3090. K IS THE LENGTH OF THE VIDEO CLIP.

Method	K	FPS	F-mAP (%)	V-mAP (%)	GFLOPs	Params
YOWO	16	34	80.4	48.8	43.7	121.4 M
YOWOv2-T	16	50	80.5	51.3	2.9	10.9 M
YOWOv2-M	16	42	83.1	50.7	12.0	52.0 M
YOWOv2-L	16	30	85.2	52.0	53.6	109.7 M
YOWOv2-T	32	50	83.0	51.2	4.5	10.9 M
YOWOv2-M	32	40	83.7	52.5	12.7	52.0 M
YOWOv2-L	32	22	87.0	52.8	91.9	109.7 M

and the results are summarized in Table I. The table shows that the decoupled fusion head outperforms the coupled fusion head. These results indicate that feature fusion should be performed decoupled due to the semantic differences between categorical and regressive features. Although the decoupled fusion head slightly slows down the detection speed, the significant improvement in performance compensates for the marginal loss in speed.

#### D. Effectiveness of the loss balance factor

We also verify the effect of loss balance factor  $\lambda$  defined in Eq.(7). Table II summarizes the results on the UCF101-24 and AVA. From the table, YOWOv2 is insensitive to  $\lambda$  in the range 3 to 6, but the larger or smaller  $\lambda$  weakens the performance of the YOWOv2. Therefore, we set  $\lambda$  to 5 in the experiments.

#### E. Comparison with YOWO

To compare the accuracy, speed, and computation of YOWOv2 with YOWO [13], we design three scales

TABLE IV  
COMPARISON WITH YOWO ON THE AVA. FPS IS MEASURED ON A GPU RTX 3090.

Method	K	FPS	mAP	GFLOPs
YOWO	16	31	17.9	44
YOWO	32	23	19.1	82
YOWO+LFB	-	-	20.2	-
YOWOv2-T	16	49	14.9	3
YOWOv2-M	16	41	18.4	12
YOWOv2-L	16	29	20.2	54
YOWOv2-T	32	49	15.6	5
YOWOv2-M	32	40	18.4	13
YOWOv2-L	32	22	21.7	92

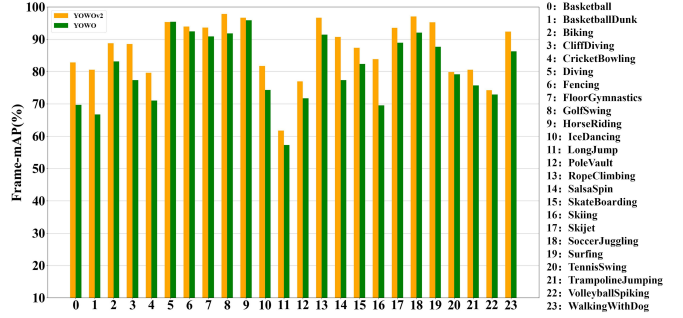


Fig. 4. Per-class frame mAP at 0.5 IoU on the UCF101-24. The orange bars represent the per-class AP of YOWOv2-L, while the green bars represent the per-class AP of YOWO.

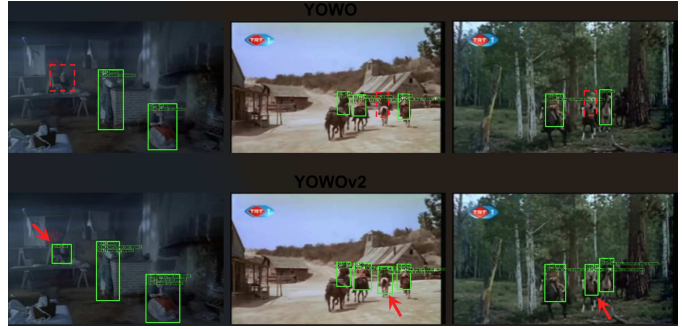


Fig. 5. Performance comparison of small action instance detection between YOWO and YOWOv2 on the AVA.

of YOWOv2 by combining different 3D backbones and 2D backbones: YOWOv2-Tiny (YOWOv2-T), YOWOv2-Medium (YOWOv2-M), and YOWOv2-Large (YOWOv2-L). To demonstrate the superior speed and performance balance of YOWOv2, we compare our YOWOv2 family with YOWO on the UCF101-24. The comparison results are summarized in Table III. The table shows that YOWOv2-T outperforms YOWO in terms of both frame mAP (80.5 % v.s. 80.4 %) and video mAP (51.3 % v.s. 48.8 %) with significantly fewer FLOPs (2.9G vs. 43.7G) and parameters (10.9M vs. 121.4M), while achieving higher FPS (50 vs. 34) on an RTX 3090 GPU. Moreover, with a stronger 2D backbone and 3D backbone, YOWOv2-L achieves the best performance. Fig. 4 shows the per-class AP comparison results between YOWO and YOWOv2-L.

We also conduct a comparative experiment with YOWO on the AVA benchmark. The comparison results are summarised in Table IV. It is unrealistic to expect the little YOWOv2-T to outperform YOWO, which has greater calculations and more parameters, given that the AVA is a highly difficult dataset. Here, the YOWOv2-L is what we focus on most. YOWOv2-L accomplishes a superior trade-off between performance and detection speed as compared to YOWO. YOWOv2-L performs better than YOWO with the LFB as well. The benefits on these well-known benchmarks demonstrate that YOWOv2's design is superior to that of YOWO, meeting the goals of inheritance and development and generating a new generation of real-time action detection framework.

On the other hand, to illustrate the effectiveness of

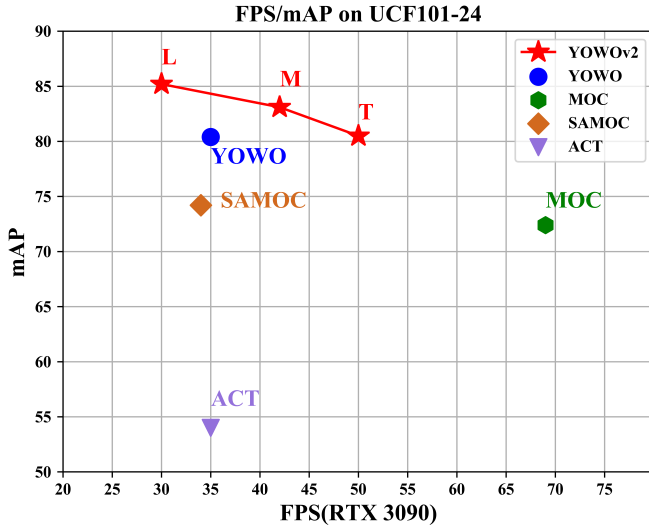


Fig. 6. Speed/accuracy trade-off among multiple real-time action detectors, including YOWO, MOC, SAMOC, ACT and the proposed YOWOv2. Speed is measured on an NVIDIA 3090 GPU with batch size 1. Note that the length of the input video clip is 16 YOWO and YOWOv2.

YOWOv2’s multi-level detection, we compare the performance of small action instance detection with YOWO, as shown in Fig. 5. The figure shows that YOWO misses certain smaller action instances because it has insufficient confidence in them (red dotted line boxes). YOWOv2 can more effectively detect the tiny action instances that YOWO cannot handle since it is equipped with the multi-level detection pipeline.

#### F. Comparison with other real-time action detectors

We contrast YOWOv2 with additional real-time action detectors in addition to YOWO. Fig. 6 shows the speed/accuracy trade-off of those detectors that can run at over 25 FPS on an RTX 3090 GPU, including YOWO [13], MOC [10], SAMOC [24] and ACT [8]. As shown in the figure, YOWOv2 greatly improves on the performance and detection speed trade-off. YOWOv2 can be seen as a new generation of superior real-time motion detectors as a result.

#### G. Comparison with state-of-the-art works

**UCF101-24.** Table V summarizes the comparison results with state-of-the-art works on the UCF101-24. For stronger performance, we also use  $K = 32$  to train and test YOWOv2. The majority of 2D CNN-based detectors extract richer spatio-temporal characteristics from the optical flow in parallel with the video clip to improve their performance. Unfortunately, using optical flow not only reduces the model’s applicability because it’s challenging to get high-quality optical flow online in real-time, but it also slows down the speed of detection. Our real-time YOWOv2 still performs admirably when measured against the potent 3D CNN-based techniques.

**AVA.** Table VI summarizes the comparison results on the AVA. Since the AVA is a very challenging benchmark where the data scene is changeable, and each action instance is labeled with multiple annotations, most current works take

TABLE V  
COMPARISON WITH STATE-OF-THE-ART WORKS ON THE UCF101-24. WE REPORT FRAME MAP AT 0.5 IOU AND VIDEO MAP AT 0.5 IOU ON THE FIRST SPLIT.

	Method	RGB	Flow	F-mAP (%)	V-mAP (%)
3D	T-CNN [19]	✓	✗	41.4	-
	I3D [18]	✓	✓	76.6	<b>59.9</b>
	Tuber [5]	✓	✗	83.2	58.4
2D	ACT [8]	✓	✓	67.1	51.4
	TACNet [9]	✓	✓	72.1	54.4
	MOC [10]	✓	✗	73.1	51.0
	MOC [10]	✓	✓	78.0	53.8
	SAMOC [24]	✓	✗	74.2	49.8
	SAMOC [24]	✓	✓	79.3	52.5
	YOWOv2-T	✓	✗	80.5	51.3
	YOWOv2-M	✓	✗	83.1	50.7
	YOWOv2-L	✓	✗	85.2	52.0
	YOWOv2-T (K=32)	✓	✗	83.0	51.2
	YOWOv2-M (K=32)	✓	✗	83.7	52.5
	YOWOv2-L (K=32)	✓	✗	<b>87.0</b>	52.8

advantage of the 3D CNN to challenge this dataset. Since the FLOPs of these 3D CNN-based detectors are too high to run in real-time, we attribute them to *Non real-time spatio-temporal action detectors*.

From the table, Tuber is a state-of-the-art action detector with the highest mAP on the AVA. However, its GFLOPs is as high as 120 and detection speed is as low as 3 FPS, although it achieves 31.7 % mAP. While having excellent performance, such a sluggish detector is exceedingly difficult to use in practical situations. Contrary to Tuber, we currently place more emphasis on the practicalities, specifically the GFLOPs and the FPS, despite the fact that the mAP metric is quite vital. Although YOWOv2 has a lower mAP than Tuber, its detection speed can satisfy real-time requirements (over 20 FPS), making it possible to use it in real-world situations to complete tasks.

Fig. 7 shows some qualitative results on the AVA. From the figure, we can see that YOWOv2 can accurately detect the basic postures, such as **walk** and **stand** and other actions of each person. This result shows that YOWOv2 has the ability to understand multiple behaviors that occur in a person, which is helpful for in-depth understanding of human behavioral intentions in the future.

#### H. Test in real scenarios

To demonstrate the generalization of YOWOv2, we also test the performance of YOWOv2 in real scenarios. Fig. 8 shows a demo of YOWOv2 in a real scene. The input frame is reshaped to  $224 \times 224$ , following the requirements in Sec. IV-B. Since most of the atomic actions in the AVA dataset do not appear in our real scenes, we only show fourteen basic action poses [18] in Fig. 8, including bend or bow, crawl, stand, walk, sit, etc. From the figure, we can see that YOWOv2 still works well in real scenarios, demonstrating its effectiveness and generalization.

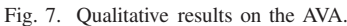
## V. CONCLUSION

In this paper, we propose a novel real-time detection framework YOWOv2 for spatial-temporal action detection.

TABLE VI

Non real-time spatio-temporal action detector

Real-time spatio-temporal action detector



to fuse multi-level features from both the 3D backbone and 2D backbone, not just the 2D backbone.

- [1] A. Clapés, À. Pardo, O. Pujol Vila, and S. Escalera, “Action detection fusing multiple kinects and a wimu: An application to in-home assistive technology for the elderly,” *Machine Vision and Applications*, vol. 29, no. 5, pp. 765–788, 2018.
- [2] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, “Stat: Spatial-temporal attention mechanism for video captioning,” *IEEE transactions on multimedia*, vol. 22, no. 1, pp. 229–241, 2019.
- [3] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, “Video action transformer network,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 244–253, 2019.
- [4] J. Wu, Z. Kuang, L. Wang, W. Zhang, and G. Wu, “Context-aware rcnn: A baseline for action detection in videos,” in *European Conference on Computer Vision*, pp. 440–456, Springer, 2020.

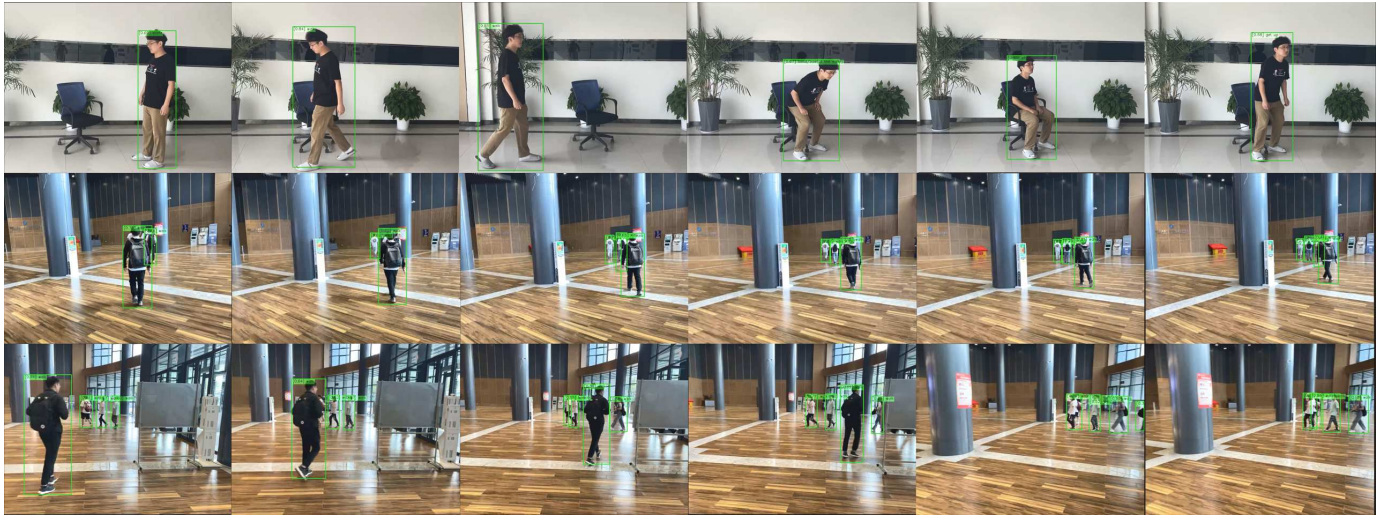


Fig. 8. Qualitative results on real scene. Since most of the atomic actions in the AVA dataset do not appear in our real scenes, we only show fourteen basic action poses, including bend or bow, crawl, stand, walk, sit, etc. The green bounding box represents the spatial localization. The action category with confidence score is shown in the upper left corner of the bounding box.

- [5] J. Zhao, Y. Zhang, X. Li, H. Chen, B. Shuai, M. Xu, C. Liu, K. Kundu, Y. Xiong, D. Modolo, *et al.*, “Tuber: Tubelet transformer for video action detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13598–13607, 2022.
- [6] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.
- [7] C. Feichtenhofer, H. Fan, J. Malik, and K. He, “Slowfast networks for video recognition,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.
- [8] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, “Action tubelet detector for spatio-temporal action localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4405–4413, 2017.
- [9] L. Song, S. Zhang, G. Yu, and H. Sun, “Tacnet: Transition-aware context network for spatio-temporal action detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11987–11995, 2019.
- [10] Y. Li, Z. Wang, L. Wang, and G. Wu, “Actions as moving points,” in *European Conference on Computer Vision*, pp. 68–84, Springer, 2020.
- [11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European conference on computer vision*, pp. 21–37, Springer, 2016.
- [12] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” *arXiv preprint arXiv:1904.07850*, 2019.
- [13] O. Köpükü, X. Wei, and G. Rigoll, “You only watch once: A unified cnn architecture for real-time spatiotemporal action localization,” *arXiv preprint arXiv:1911.06644*, 2019.
- [14] J. Redmon and A. Farhadi, “Yolo9000: better, faster, stronger,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.
- [15] O. Köpükü, N. Kose, A. Gunduz, and G. Rigoll, “Resource efficient 3d convolutional neural networks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.
- [16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.
- [17] K. Soomro, A. R. Zamir, and M. Shah, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” *arXiv preprint arXiv:1212.0402*, 2012.
- [18] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, *et al.*, “Ava: A video dataset of spatio-temporally localized atomic visual actions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6047–6056, 2018.
- [19] R. Hou, C. Chen, and M. Shah, “Tube convolutional neural network (t-cnn) for action detection in videos,” in *Proceedings of the IEEE international conference on computer vision*, pp. 5822–5831, 2017.
- [20] K. Duarte, Y. Rawat, and M. Shah, “Videocapsulenet: A simplified network for action detection,” *Advances in neural information processing systems*, vol. 31, 2018.
- [21] S. Chen, P. Sun, E. Xie, C. Ge, J. Wu, L. Ma, J. Shen, and P. Luo, “Watch only once: An end-to-end video action detection framework,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8178–8187, 2021.
- [22] J. Pan, S. Chen, M. Z. Shou, Y. Liu, J. Shao, and H. Li, “Actor-context-actor relation network for spatio-temporal action localization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 464–474, 2021.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [24] X. Ma, Z. Luo, X. Zhang, Q. Liao, X. Shen, and M. Wang, “Spatio-temporal action detector with self-attention,” in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2021.
- [25] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” *arXiv preprint arXiv:2207.02696*, 2022.
- [26] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “Yolox: Exceeding yolo series in 2021,” *arXiv preprint arXiv:2107.08430*, 2021.
- [27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*, pp. 740–755, Springer, 2014.
- [28] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3146–3154, 2019.
- [29] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, 2019.
- [30] L. Sui, C.-L. Zhang, L. Gu, and F. Han, “A simple and efficient pipeline to build an end-to-end spatial-temporal action detector,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5999–6008, 2023.