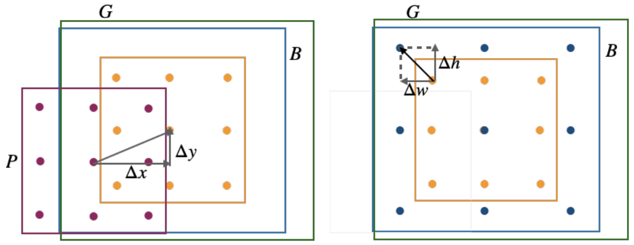
Spatial Feature Calibration and Temporal Fusion for Effective One-stage Video Instance Segmentation (Supplemental Materials)

Minghan Li[1,2], **Shuai Li**[1,2], **Lida Li**[1] and **Lei Zhang**[1,2*]
[1]The HongKong Polytechnic University, [2]DAMO Academy, Alibaba Group
liminghan0330@gmail.com, { csshuaili, cslli, cslzhang}@comp.polyu.edu.hk

1. Aligned Features on Bounding Boxes



(a) Scale-invariant Translation (b) Scale Transformation

Figure 1. Illustration of aligned features from anchor boxes to predicted bounding boxes.

The derivation process of generating the offsets for deformable convolution can also be divided into two steps: *scale-invariant translation* and *scale transformation*.

Scale-invariant translation: Fig. 1 (a) demonstrates the scale-invariant translation on the sampling locations of a 3×3 standard convolution. All sampling points on the grid $\mathcal{R}$ of the anchor box have the same scale-invariant translation as the centre point. Based on the first two functions of the regression transformation d, the scale-invariant translation of sampling points can be calculated as $P_w d_x$ and $P_h d_y$ on x and y axis respectively. It is need to normalise the scale-invariant translations on the grid $\mathcal{R}$ to get its corresponding offsets on x and y axis:

$$\Delta x = \frac{P_w d_x}{1/k_w P_w} = k_w d_x, \quad \Delta y = \frac{P_h d_y}{1/k_h P_h} = k_h d_y, \quad (1)$$

where $1/k_w P_w$ and $1/k_h P_h$ are the normalised units of the grid $\mathcal{R}$ on the x and y axes, $k = (k_h, k_w)$ is the kernel size of the height and the width. After the scale-invariant translation, the grid $\mathcal{R}$ will be translated to the yellow grid $\mathcal{R}_t$ of Fig. 1 (a):

$$\mathcal{R}_t = \mathcal{R} + (\Delta y, \Delta x)I, \quad (2)$$

where I is a matrix with all elements as 1.

Scale transformation: As shown in Fig. 1 (b), according to the last two functions of the regression transformation, a scale transformation also needs to be implemented on the grid $\mathcal{R}_t$ to align convolutional grid with regressed bounding boxes. Based on the log-space translations of the width and height of anchor box d_w, d_h, the absolute scale offsets of the width and the height on the grid $\mathcal{R}_t$ should be

$$\Delta h = (B_h/P_h) - 1 = \exp(d_h) - 1, \quad (3)$$
$$\Delta w = (B_w/P_w) - 1 = \exp(d_w) - 1. \quad (4)$$

Note that the scale transformations of all points on the grid $\mathcal{R}_t$ are closely related to the grid position. Specifically, the scaled grid $\mathcal{R}_{t+s}$ (blue points in Fig. 1 (b)) should be

$$\begin{aligned} \mathcal{R}_{t,s} &= \mathcal{R}_t + (\Delta h, \Delta w)\mathcal{R} \\ &= \mathcal{R} + (\Delta y, \Delta x)I + (\Delta h, \Delta w)\mathcal{R}. \end{aligned} \quad (5)$$

where the absolute scale offsets of width and height are multiplied by the points on the grid one by one. Here, the convolutional grid $\mathcal{R}_{t+s}$ after scale-invariant translation and scale transformation has aligned with its corresponding predicted bounding box. Overall, the derived offsets for all points on the grid should be

$$\mathcal{O} = (\Delta y, \Delta x)I + (\Delta h, \Delta w)\mathcal{R}. \quad (6)$$

2. Ablation Study for Temporal Fusion Module

We first conduct experiments on different backbones and anchors to verify the validity of our proposed temporal fusion module. As shown in Table 1, the temporal fusion module brings significant performance improvement in all settings, where a deeper backbone with stronger feature representation or a model with more anchors brings greater performance gains. Besides, the last two rows of Table 1 illustrate that more anchors are not better for all backbones. Because more anchors may bring more false positives, resulting in performance degradation. To alleviate this issue, in experiments, all masks tracked from previous frames will be given a decaying confidence, ensuring that tracked masks are inferred from nearby frames to reduce the effect of false positives.

[*]Corresponding author.This work is supported by the Hong Kong RGC RIF grant (R5001-18).

Table 1. Ablation study for temporal fusion module based on different backbones and anchors. (mask AP)

Backbone	Anchors	Baseline	+TF	Improve
R-50-DCN	3	25.2	29.1	3.9
	6	26.6	29.9	3.3
	9	26.7	32.2	5.5
R-101-DCN	3	28.9	34.1	5.2
	6	30.7	37.2	6.5
	9	31.1	36.4	5.3

3. Spatial Calibration in Image Domain

Existing one-stage image instance segmentation methods cannot guarantee feature calibration on anchors. Therefore, we also perform STMask only with spatial feature calibration on COCO dataset. Experimental results shown in Table 2 verify the effectiveness of spatial calibration in image instance segmentation task as well.

Table 2. Ablation study for FCA and FCB on COCO2017 val and testdev set, where Yolact with R101-FCN backbone and Yolact++ with R101-DCN backbone).

Methods	A	FPS	val2017		testdev2017		
			B-AP	M-AP	M-AP	AP_{50}	AP_{75}
Yolact	3	33.5	32.1	29.9	29.8	48.5	31.2
+FCA	3	33.3	33.1	30.7	30.9	50.7	32.2
+FCB (ada)	3	32.8	34.8	31.6	31.9	51.9	33.5
Yolact++	9	27.3	35.6	34.5	34.6	53.8	36.9
Yolact++	3	30.8	33.9	33.0	33.2	53.2	34.8
+FCA	3	24.5	34.7	33.7	33.9	53.8	35.8
+FCB (ada)	3	24.3	36.0	34.5	34.6	54.4	36.6