Supplementary Materials for GaTector: A Unified Framework for Gaze Object Prediction

Binglu Wang*, Tao Hu, Baoshan Li, Xiaojuan Chen, Zhijie Zhang
Xi'an University of Architecture and Technology, China

1. Process to calculate the training loss.

The training process of our GaTector is driven by three loss terms, *i.e.* object detection loss $\mathcal{L}_{det}$, gaze estimation loss $\mathcal{L}_{gaze}$ and our proposed energy aggregation loss $\mathcal{L}_{eng}$. The energy aggregation loss $\mathcal{L}_{eng}$ is illustrated in SubSection 3.4 of our manuscript. Here, we elaborate on detailed processes to calculate the object detection loss $\mathcal{L}_{det}$ and the gaze estimation loss $\mathcal{L}_{gaze}$. For a fair comparison, we keep identical setups with YOLOv4 [1] when calculating the object detection loss $\mathcal{L}_{det}$, and keep identical setups with Chong *et al*. [2] when calculating the gaze estimation loss $\mathcal{L}_{gaze}$.

Object detection. Given a predicted box $(x, y, w, h, p, \mathbf{s})$, (x, y) indicates the central point, (w, h) indicates width and height, p is the predicted overlap, and $\mathbf{s} = [s_0, s_1, ..., s_C]$ is the predicted classification score. The corresponding ground truth can be represented as $(x^g, y^g, w^g, h^g, o, \mathbf{y})$, where o indicates the ground truth overlap and $\mathbf{y} = [y_0, y_1, ..., y_C]$, $y_c \in \{0, 1\}$ represents whether this bounding box belongs to the c^{th} category. The object detection loss $\mathcal{L}_{det}$ jointly considers classification, overlapping, and box regression.

$$\mathcal{L}_{det} = \mathcal{L}_{det}^{cls} + \mathcal{L}_{det}^{o} + \mathcal{L}_{det}^{reg}. \tag{1}$$

The classification term calculates the binary cross-entropy loss:

$$\mathcal{L}_{det}^{cls} = \frac{1}{C+1}\sum_{c=0}^{C} -[y_c \log \hat{s}_c + (1-y_c)\log(1-\hat{s}_c)], \tag{2}$$

where $\hat{s}_c$ indicates the classification score after the sigmoid activation.

The overlapping loss adopts the binary cross-entropy loss as well:

$$\mathcal{L}_{det}^{o} = -[o \log \hat{p} + (1-o)\log(1-\hat{p})], \tag{3}$$

where $\hat{p}$ indicates the overlap score after the sigmoid activation.

As for box regression, we adopt the CIoU loss [3]. Succinctly, the regression loss can be calculated as follows:

$$\mathcal{L}_{det}^{reg} = 1 - IoU + \frac{\rho^2\left((x,y),(x^g,y^g)\right)}{d^2} + \alpha v, \tag{4}$$

where IoU indicates the intersection over union between the predicted box and the ground truth box. The term $\frac{\rho^2((x,y),(x^g,y^g))}{d^2}$ aims to minimize the distances between central points of two boxes, where $\rho^2\left((x,y),(x^g,y^g)\right)$ indicates the Euclidean distance and d represents the diagonal length of the smallest enclosing box covering the two boxes. In addition, the term αv measures the consistency of aspect ratio, where v is defined as:

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^g}{h^g} - \arctan\frac{w}{h}\right)^2. \tag{5}$$

The coefficient α can be calculated as:

$$\alpha = \frac{v}{(1-IoU)+v}. \tag{6}$$

Please refer to [3] for more details about CIoU loss.

Gaze estimation. We follow Chong *et al*. [2] to calculate the gaze estimation loss $\mathcal{L}_{gaze}$. Given the annotated gaze point $\mathbf{q} = (q_x, q_y)$, we apply the Gaussian blur to generate the vanilla ground truth heatmap $\mathbf{T}'$.

$$\mathbf{T}' = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{1}{2}\left(\frac{(x-q_x)^2}{\sigma_x^2} + \frac{(y-q_y)^2}{\sigma_y^2}\right)\right]. \tag{7}$$

In Eq.(7), σ_x and σ_y indicate the standard deviation. We follow Chong *et al*. [2] and set $\sigma_x = 3$, $\sigma_y = 3$. Afterward, we normalize the heatmap and obtain the ground truth heatmap $\mathbf{T} = \mathbf{T}/\max(\mathbf{T}')$. Given a predicted heatmap $\mathbf{M} \in \mathbb{R}^{H\times W}$, we calculate the mean square error to obtain the gaze estimation loss $\mathcal{L}_{gaze}$:

$$\mathcal{L}_{gaze} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}(M_{i,j} - T_{i,j})^2. \tag{8}$$

*Corresponding author.

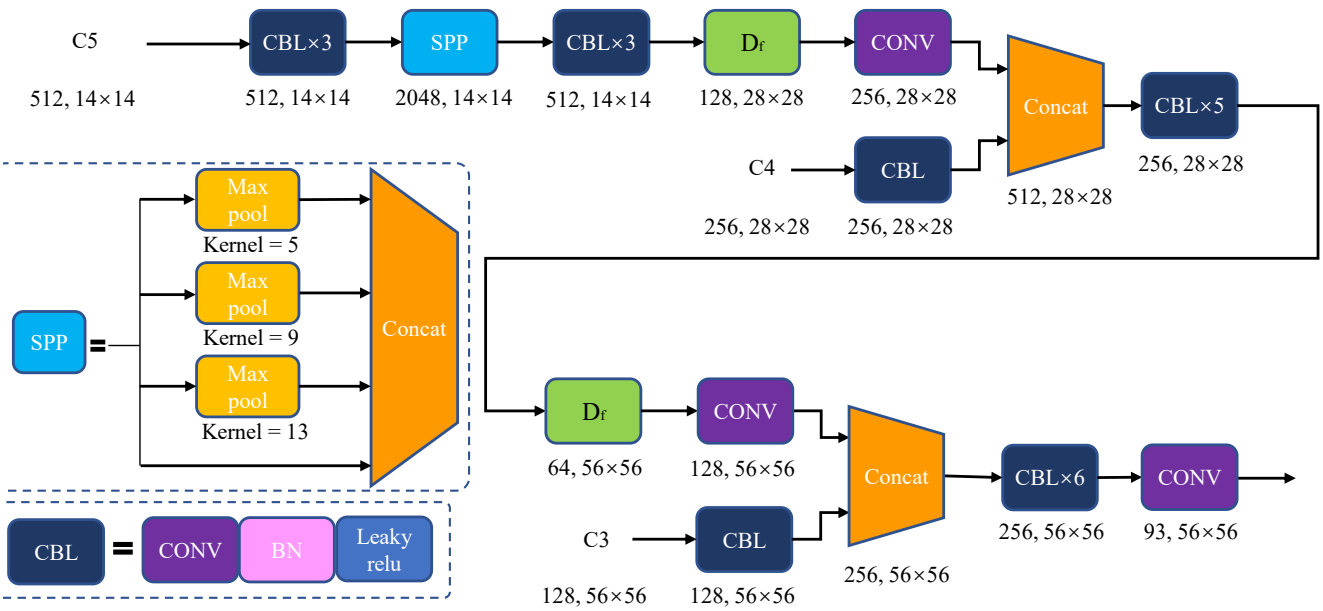

Figure 1. Detailed network architecture of our object detection branch. Df, CONV, and BN indicate the *Defcous*, convolution, and batch normalization operations, respectively. Under each operation, we present the size of the feature map in the form of (channel, height×width).

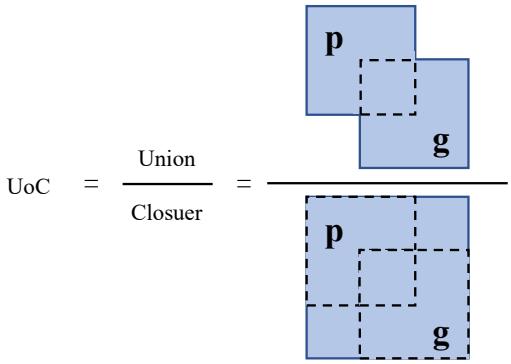

Figure 2. Calculation process of the union over closure.

2. Process to calculate the wUoC metric.

In our paper, we propose the wUoC metric to measure the performance of gaze object prediction. Given the predicted box $\mathbf{p}$ and the ground truth box $\mathbf{g}$, we calculate their minimum closure and obtain a bounding box $\mathbf{a}$. Then, Figure 2 illustrates the process to calculate the UoC (union over closure).

Afterward, we further introduce a size similarity weight into the UoC metric. The size similarity weight considers the area of two boxes and can be defined as $\min(\frac{\mathbf{p}}{\mathbf{g}}, \frac{\mathbf{g}}{\mathbf{p}})$. Thus, our proposed metric can be formulated as:

$$\text{wUoC} = \min(\frac{\mathbf{p}}{\mathbf{g}}, \frac{\mathbf{g}}{\mathbf{p}}) \times \frac{\mathbf{p} \cup \mathbf{g}}{\mathbf{a}}. \tag{9}$$

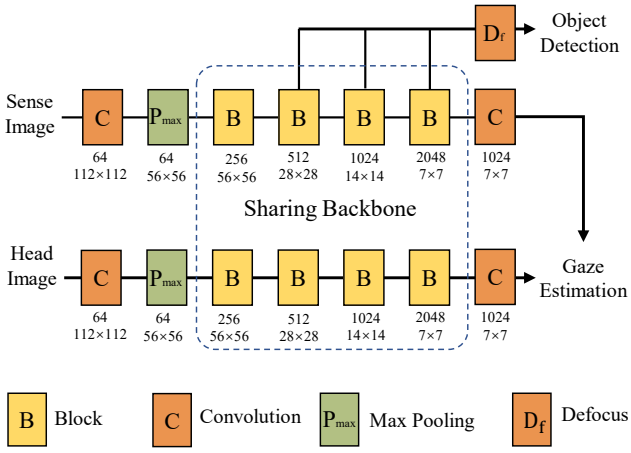

Figure 3. Detailed network architecture of the specific-general-specific feature extractor. Under each operation, we present the size of the feature map in the form of (channel, height×width).

3. Detailed architecture of each component

Figure 3 illustrates the detailed architecture of our backbone network. Given a sense image and a head image, we first employ two specific convolutional layers to convert these two inputs into the general space. Then, a sharing backbone network with four blocks processes two inputs and generates features. Finally, we select features from the last three backbone layers (*i.e.* C_3, C_4, C_5), use the *Defcous* layer to enlarge the feature map, and prepare specific inputs for the object detection branch. Simultaneously, we utilize

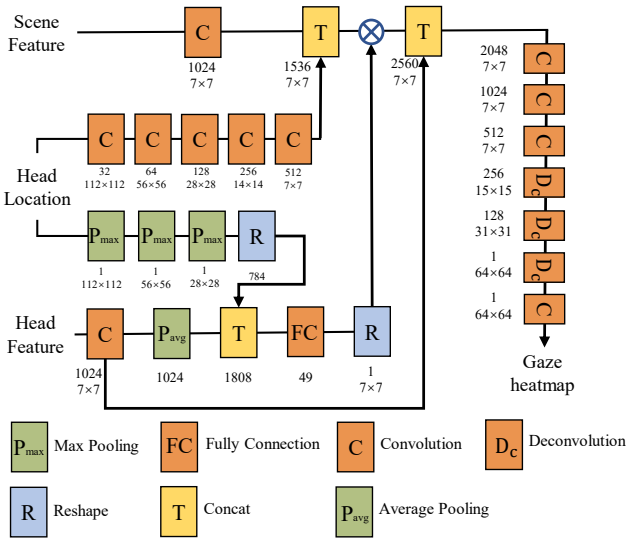

Figure 4. Detailed network architecture of the gaze estimation branch. Under each operation, we present the size of the feature map in the form of (channel, height×width).

two convolutional layers to prepare inputs for the gaze estimation branch.

Figure 1 presents the detection branch. There are three inputs with different sizes, *i.e.* C_5, C_4, C_3, which are gradually integrated to detect objects.

Figure 4 presents the detailed process to estimate the gaze heatmap. Given the head location map, we employ five convolutional layers to extract the location feature, which is concatenated with the sense feature. Simultaneously, we jointly consider head location and head feature to predict an attention map, which is used to modulate the fused feature. Afterward, we employ three convolutional layers to abstract features, use three deconvolutional layers to enlarge the feature map, and utilize a convolutional layer to estimate the gaze heatmap.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 1

[2] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *CVPR*, pages 5396–5406, 2020. 1

[3] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *AAAI*, volume 34, pages 12993–13000, 2020. 1