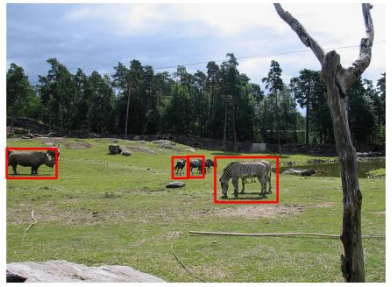
Focal and Global Knowledge Distillation for Detectors

Zhendong Yang[*1,2] Zhe Li[2] Xiaohu Jiang[1] Yuan Gong[1]
Zehuan Yuan[2] Danpei Zhao[3] Chun Yuan[†1]
[1]Tsinghua Shenzhen International Graduate School [2]ByteDance Inc
[3]BeiHang University

CVPR 2022

Introduce

knowledge distillation + object detection



- foreground-background class imbalance

  Distill the whole feature may introduce much noise because of the imbalance between the foreground and background.

- only focus on foreground

  Considering that background regions are useless or even harmful for distillation.

	distillation area			mAP	mAR
	fg	bg	split		
RetinaNet Res101-Res50	×	×		37.4	53.9
	✓	×		39.3	55.6
	×	✓		39.2	55.8
	✓	✓	×	38.9	55.1
	✓	✓	✓	**39.4**	**56.1**

Introduce

- The difference between student’ s attention and teacher’ s attention in the foreground is quite significant.

  ➡ **focal distillation**

  focal distillation calculates the attention of different pixels and channels in teacher’ s feature, allowing the student to focus on teacher’ s crucial pixels and channels.

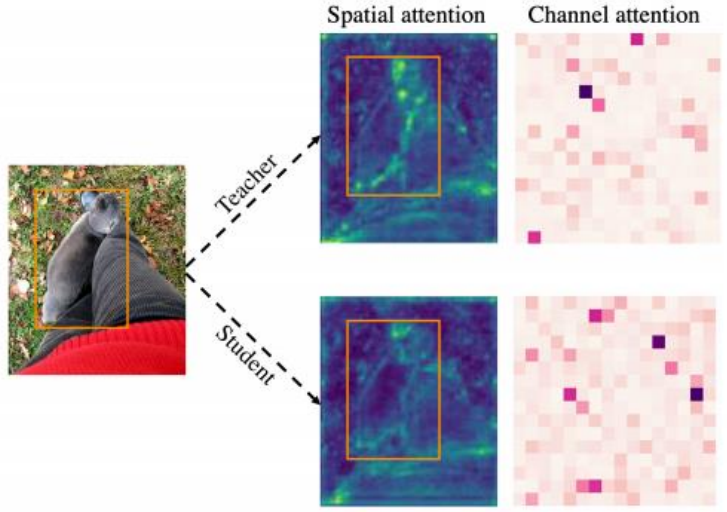


Figure 1. Visualization of the spatial and channel attention map from the teacher detector (RetinaNet-ResNeXt101) and the student detector (RetinaNet-ResNet50).

- It is generally acknowledged that the relation between different objects contains valuable information in object detection.

  ➡ **global distillation**

  we utilize GcBlock to extract the relation between different pixels and then distill them from teachers to students.

focal distillation + global distillation

Method

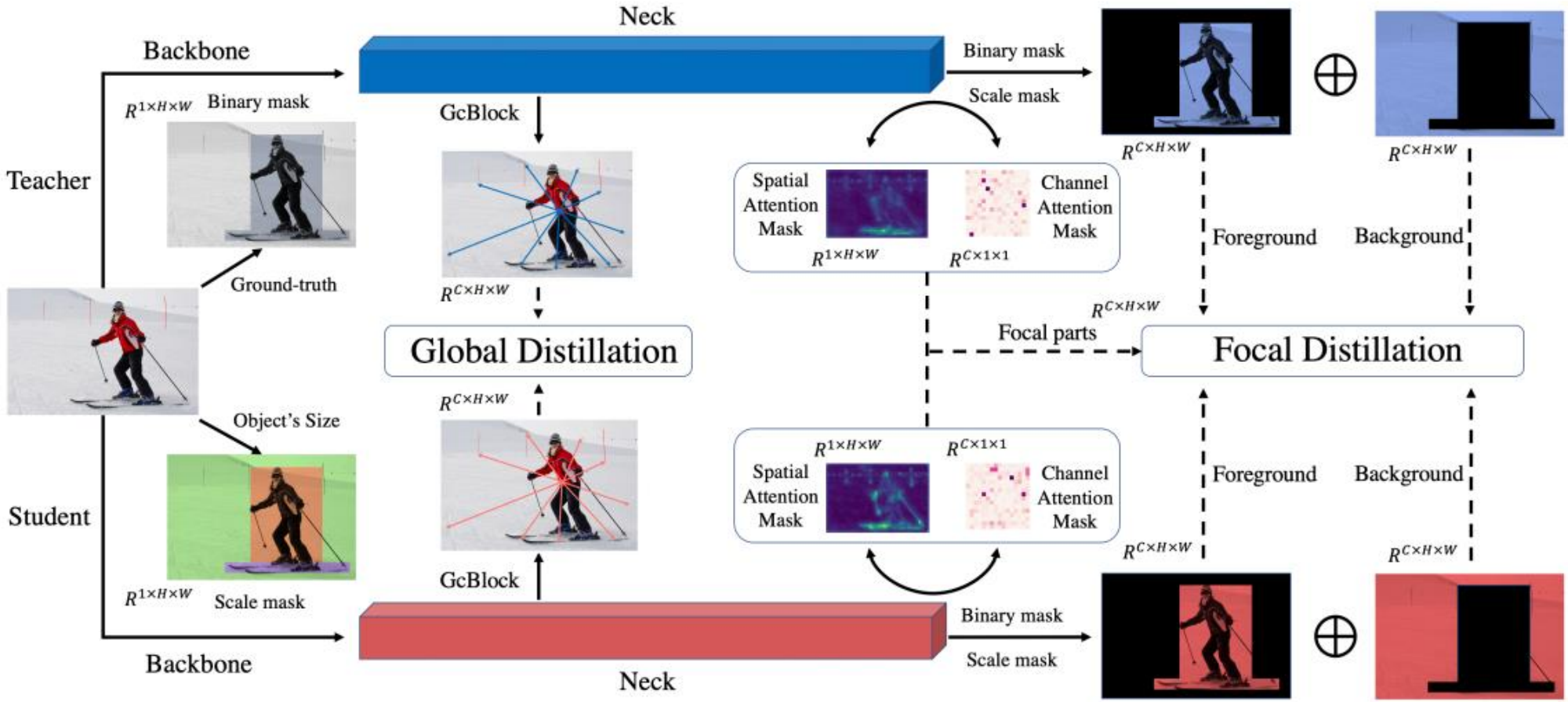


Figure 2. An illustration of FGD, including focal distillation and global distillation. Focal distillation not only separates the foreground and the background, but also enables the student network to better pay attention to the important information in the teacher network's feature map. Global distillation bridges the gap between the global context of the student and the teacher.

Focal Distillation

binary mask $$M_{i,j} = \begin{cases} 1, & \text{if } (i,j) \in r \\ 0, & \text{Otherwise} \end{cases}$$

scale mask $$S_{i,j} = \begin{cases} \frac{1}{H_r W_r}, & \text{if } (i,j) \in r \\ \frac{1}{N_{bg}}, & \text{Otherwise} \end{cases} \qquad N_{bg} = \sum_{i=1}^{H} \sum_{j=1}^{W} (1 - M_{i,j})$$

spatial attention mask $$A^S(F) = H \cdot W \cdot softmax\big(G^S(F)/T\big) \qquad G^S(F) = \frac{1}{C} \cdot \sum_{c=1}^{C} |F_c|$$

channel attention mask $$A^C(F) = C \cdot softmax\big(G^C(F)/T\big) \qquad G^C(F) = \frac{1}{HW} \cdot \sum_{i=1}^{H} \sum_{j=1}^{W} |F_{i,j}|$$

Focal Distillation

distillation loss $$L_{fea} = \frac{1}{CHW} \sum_{k=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(F_{k,i,j}^{T} - f(F_{k,i,j}^{S}) \right)^2$$

feature loss $$L_{fea} = \alpha \sum_{k=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} M_{i,j} S_{i,j} A_{i,j}^{S} A_{k}^{C} \left(F_{k,i,j}^{T} - f(F_{k,i,j}^{S}) \right)^2$$
$$+\beta \sum_{k=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} (1 - M_{i,j}) S_{i,j} A_{i,j}^{S} A_{k}^{C} \left(F_{k,i,j}^{T} - f(F_{k,i,j}^{S}) \right)^2$$

attention loss $$L_{at} = \gamma \cdot \left(l(A_t^S, A_S^S) + l(A_t^C, A_S^C) \right)$$

focal loss $$L_{focal} = L_{fea} + L_{at}$$

Global Distillation

global loss $$L_{global} = \lambda \cdot \sum \left(\mathcal{R}(F^T) - \mathcal{R}(F^S) \right)^2$$

$$\mathcal{R}(F) = F + W_{v2}(ReLU(LN(W_{v1}(\sum_{j=1}^{N_p} \frac{e^{W_k F_j}}{\sum_{m=1}^{N_p} e^{W_k F_M}} F_j))))$$

Wk, Wv1 and Wv2 denote convolutional layers, LN denotes the layer normalization, Np is the number of pixels in the feature and λ is a hyper-parameter to balance the loss.

overall loss $$L = L_{original} + L_{focal} + L_{global}$$

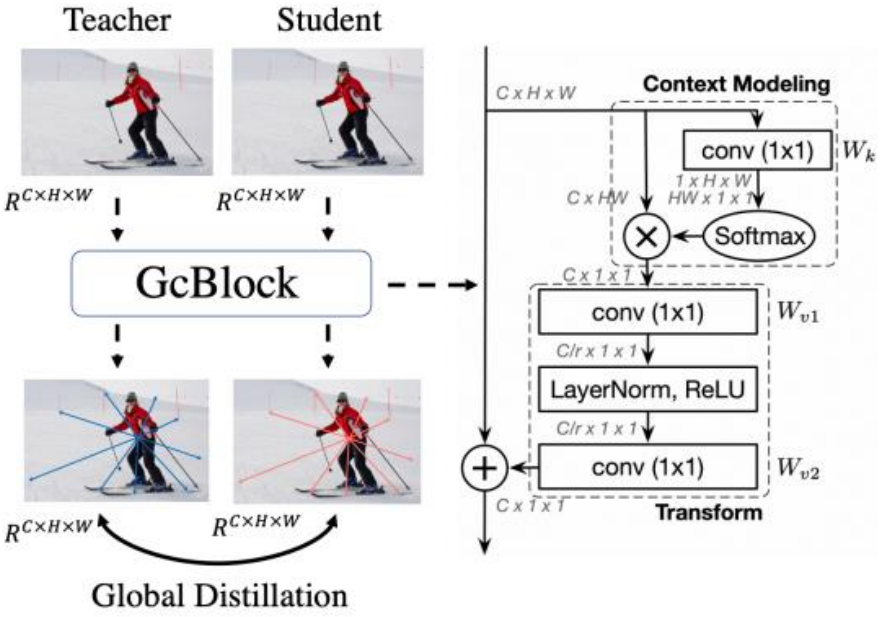


Figure 4. The Global Distillation with GcBlock. The inputs are the feature maps from the teacher's neck and student's neck, respectively.

Experiments

Method	mAP	AP_S	AP_M	AP_L
RetinaNet-Res101(T)	38.9	21.0	42.8	52.4
RetinaNet-Res50(S)	37.4	20.6	40.7	49.7
FGFI [31]	38.6	21.4	42.5	51.5
GID [6]	39.1	22.8	43.1	52.3
Ours	39.6	22.9	43.7	53.6
Ours †	39.7	22.0	43.7	53.6
RCNN-Res101(T)	39.8	22.5	43.6	52.8
RCNN-Res50(S)	38.4	21.5	42.1	50.3
FGFI [31]	39.3	22.5	42.3	52.2
GID [6]	40.2	22.7	44.0	53.2
Ours	40.4	22.8	44.5	53.5
Ours †	40.5	22.6	44.7	53.2
FCOS-Res101(T)	40.8	24.2	44.3	52.4
FCOS-Res50(S)	38.5	21.9	42.8	48.6
GID [6]	42.0	25.6	45.8	54.2
Ours	42.1	27.0	46.0	54.6
Ours †	42.7	27.2	46.5	55.5

Table 2. Results of different distillation methods with different detection frameworks on COCO dataset. **T** and **S** mean the teacher and student detector, respectively. FGFI can only be applied to an anchor-based detector. † means using inheriting strategy. We train the FCOS with tricks including GIoULoss, norm-on-bbox and center-sampling which is the same as GID.

Teacher	Student	mAP	AP_S	AP_M	AP_L	mAR	AR_S	AR_M	AR_L
RetinaNet ResNeXt101	RetinaNet-Res50	37.4	20.6	40.7	49.7	53.9	33.1	57.7	70.2
	FKD [39]	39.6(+2.2)	22.7	43.3	52.5	56.1(+2.2)	36.8	60.0	72.1
	Ours	40.4(+3.0)	23.4	44.7	54.1	56.7(+2.8)	37.6	61.5	72.4
	Ours†	40.7(+3.3)	22.9	45.0	54.7	56.8(+2.9)	36.5	61.4	72.8
Cascade Mask RCNN ResNeXt101	Faster RCNN-Res50	38.4	21.5	42.1	50.3	52.0	32.6	55.8	66.1
	FKD [39]	41.5(+3.1)	23.5	45.0	55.3	54.4(+2.4)	34.0	58.2	69.9
	Ours	42.0(+3.6)	23.8	46.4	55.5	55.4(+3.4)	35.5	60.0	70.0
RepPoints ResNeXt101	RepPoints-Res50	38.6	22.5	42.2	50.4	55.1	34.9	59.4	70.3
	FKD [39]	40.6(+2.0)	23.4	44.6	53.0	56.9(+1.8)	37.3	60.9	71.4
	Ours	41.3(+2.7)	24.5	45.2	54.0	58.4(+3.3)	39.1	62.9	74.2
	Ours†	42.0(+3.4)	24.0	45.7	55.6	58.2(+3.1)	37.8	62.2	73.3

Table 3. Results of more detectors with stronger teacher detectors on COCO dataset. † means using inheriting strategy, which can only be applied when the student and teacher have the same head structure.

Ablation

Method	ReinaNet ResX101-Res50			
L_{focal}	-	✓	-	✓
L_{global}	-	-	✓	✓
mAP	37.4	40.2	40.2	**40.4**
AP_S	20.0	22.8	22.9	**23.4**
AP_M	40.7	44.0	44.3	**44.7**
AP_L	49.7	54.0	53.4	**54.1**
mAR	53.9	56.2	56.4	**56.7**
AR_S	33.1	36.8	37.3	**37.6**
AR_M	57.7	60.3	60.5	**61.5**
AR_L	70.2	72.3	72.2	**72.4**

Table 4. Ablation study of focal and global distillation.

Method	ReinaNet ResNeXt101-Res50			
Spatial attention	-	✓	-	✓
Channel attention	-	-	✓	✓
mAP	37.4	40.0	39.7	**40.2**
AP_S	20.0	22.3	22.0	**22.8**
AP_M	40.7	44.0	43.5	**44.0**
AP_L	49.7	53.6	53.4	**54.0**
mAR	53.9	56.1	55.8	**56.2**
AR_S	33.1	36.5	35.7	**36.8**
AR_M	57.7	60.2	59.9	**60.3**
AR_L	70.2	72.1	71.8	**72.3**

Table 5. Ablation study of the spatial and channel attention mask.

Ablation

Methods	mAP	AP_S	AP_M	AP_L
baseline	38.4	21.5	42.1	50.3
Non-Local	39.8	22.7	43.1	52.3
GcBlock	**41.5**	**23.4**	**46.0**	**55.3**

Table 6. Comparison of different global relation methods on Faster RCNN ResNeXt101-Res50. Here we train the student just with global distillation.

T	0.3	0.5	0.8	1.0	1.2
mAP	40.1	**40.4**	**40.4**	40.2	40.0
mAR	56.4	**56.7**	56.6	56.5	56.4

Table 7. Ablation study of temperature hyper-parameter T on RetinaNet ResNeXt101-Res50.

Visualization

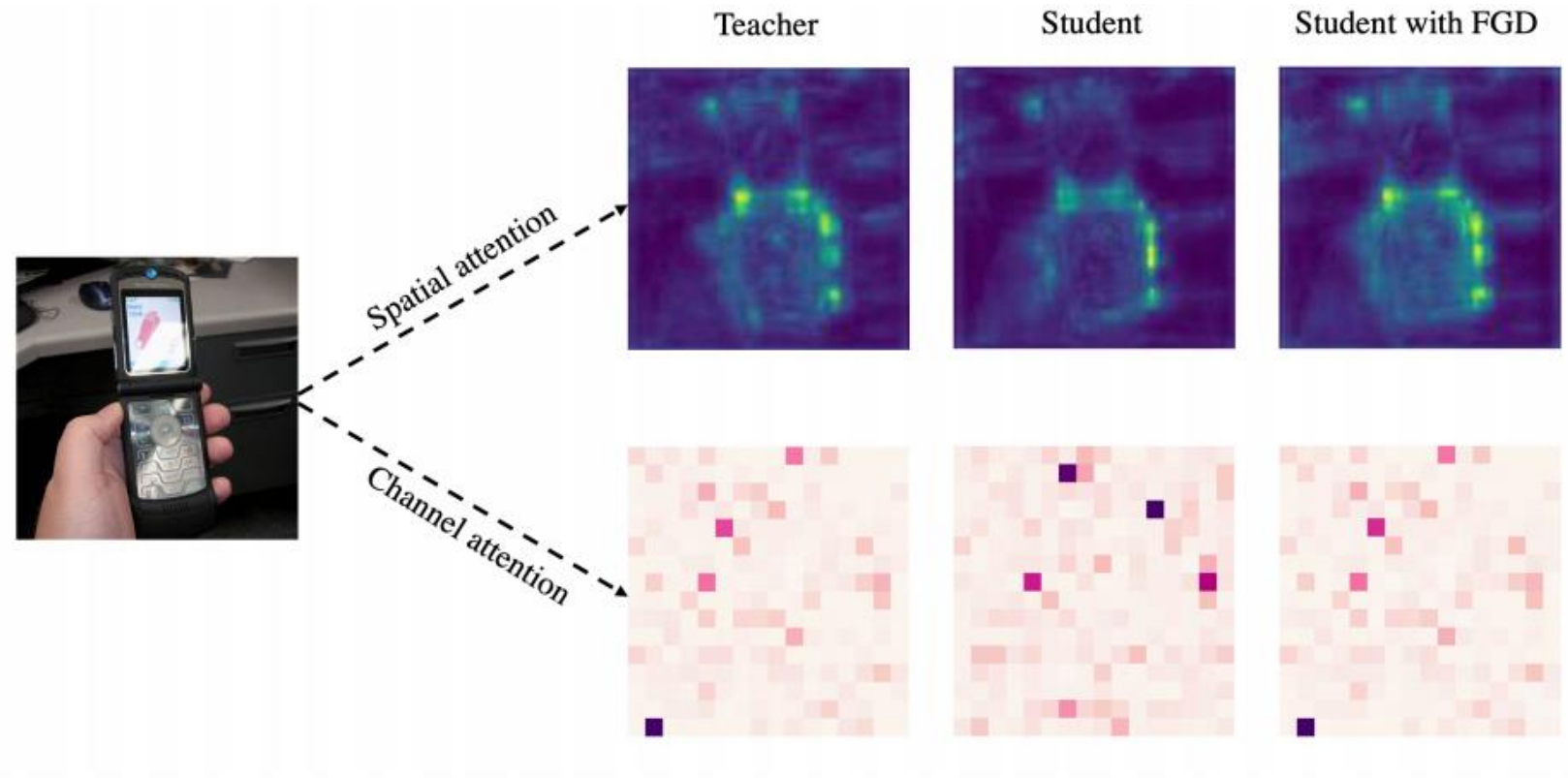


Figure 5. Visualization of the spatial and channel attention mask from different detectors. Each pixel in the channel attention mask means a channel. **Teacher detector**: RetinaNet-ResNeXt101. **Student detector**: RetinaNet-ResNet50

Thank you