DFD: Distilling the Feature Disparity Differently for Detectors

Kang Liu[1,2], Yingyi Zhang[3], Jingyun Zhang[2], Jinmin Li[2], Jun Wang[2], Shaoming Wang[2], Chun Yuan[1], Rizen Guo [2]

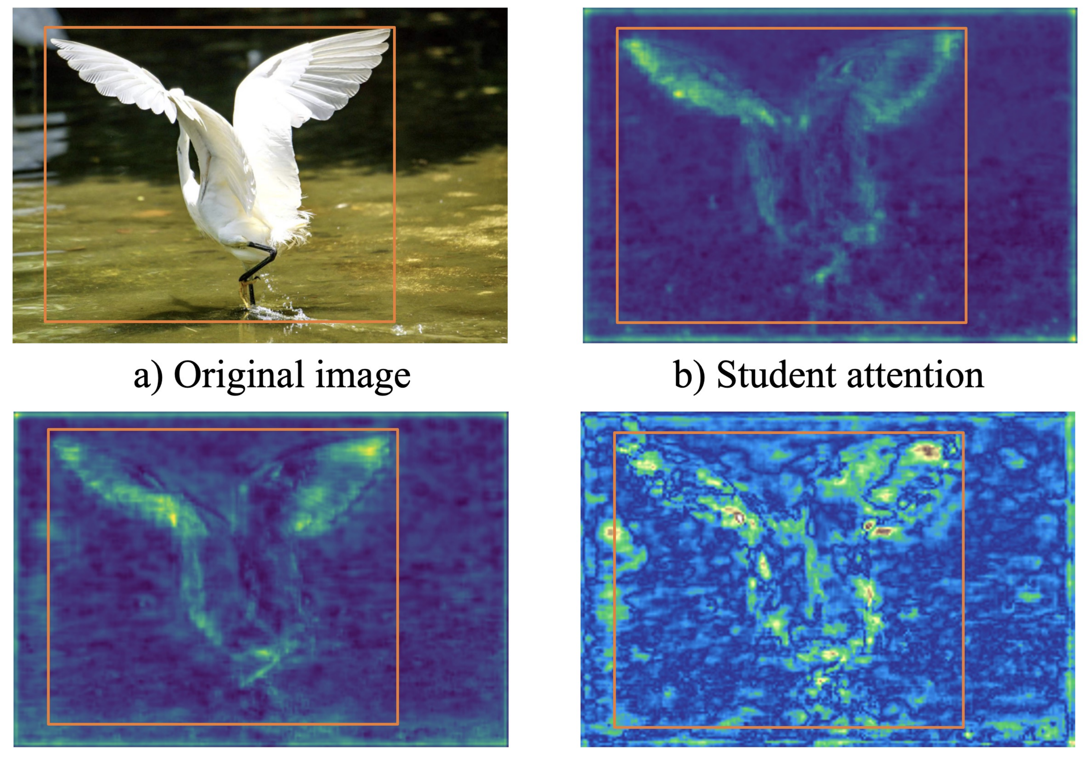
1Tsinghua University, China; 2 Tencent WeChat Pay Lab33, China; 3 Tencent Youtu Lab, China

ICML International Conference On Machine Learning

Tsinghua University 清华大学 微信支付

Motivation

Overview: *DFD* applys different treatments to regions with varying learning difficulties, simultaneously incorporating leniency and strictness, which enables better distillation efforts.



a) Original image

b) Student attention

c) Teacher attention

d) L1 distance between student and teacher

Figure 1. Visualization of spatial attention.

- **Region:** The feature responses do not fully conform to the distinction between foreground and background.
- **Disparity:** There are complex differences in both the regions and intensities of the features between the teacher and student models.

- **HD:** High disparity region
- **LD:** Low disparity region.
- **Split**: Split these regions, and use different weights for the distillation losses of different parts.
- **DC:** Using different cons-traints for different regions.

Table 1.Comparison of distillation on different regions.

Model	HD	LD	Split	DC	mAP
RetinaNet RX101 -R50	-	-	-	-	36.4
	✓	-	-	-	39.8
	-	✓	-	-	38.2
	✓	✓	-	-	39.6
	✓	✓	✓	-	40.0
	✓	✓	✓	✓	**40.4 (Ours)**

- **Knowledge Gap:** The feature disparity represents the varying capabilities of the models.
- **Distillation effort:** Using different distillation methods for different disparity regions can significantly enhance the distillation effect.

Methodology

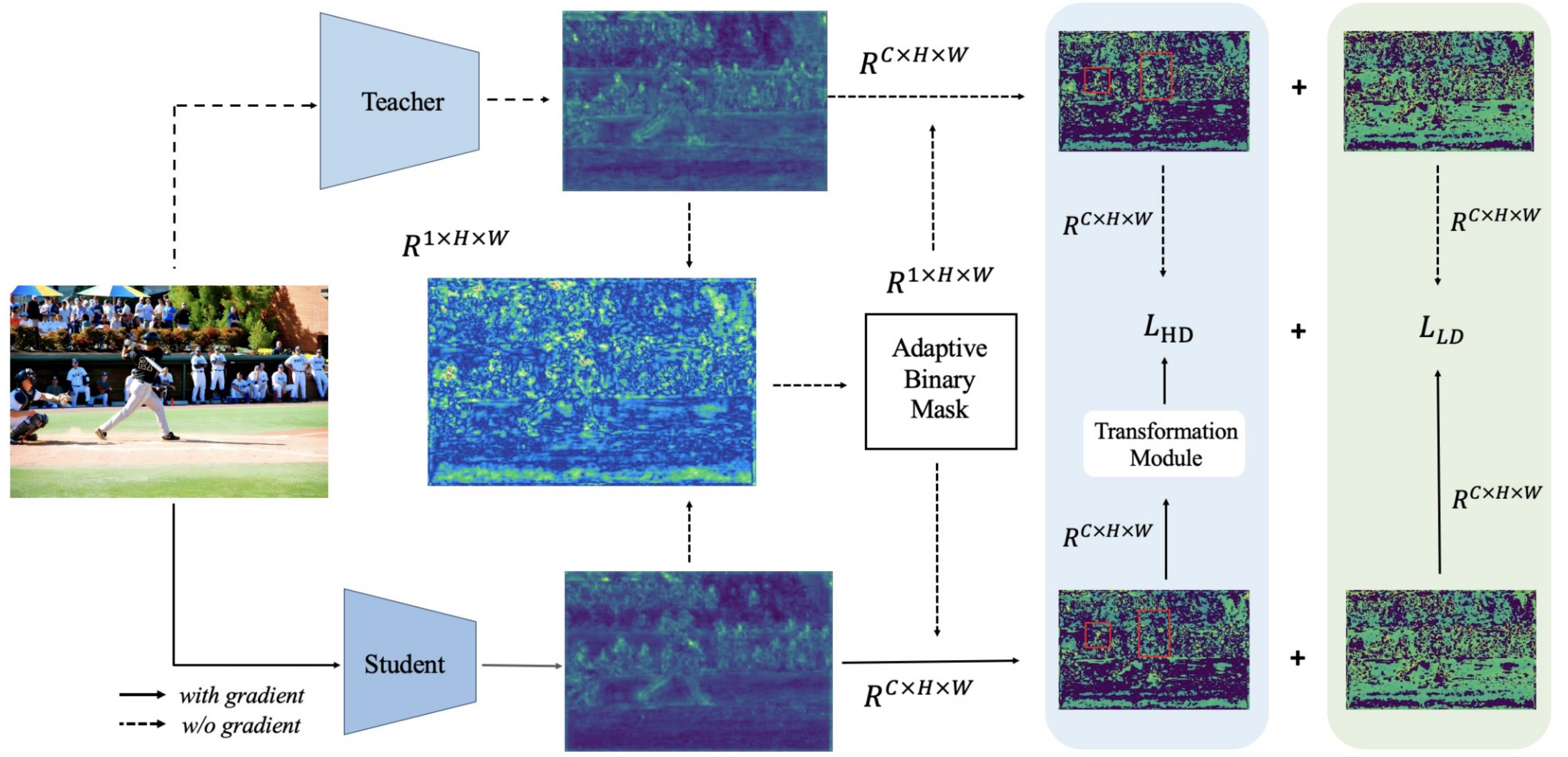


Figure 2: Overall architecture of DFD. We partition the feature into high disparity regions (HD) and low disparity regions (LD) by utilizing the mean L1 distance of spatial attention between the student and teacher as the threshold.

- Calculate the spatial attention:

$$A(F) = H \cdot W \cdot \mathrm{softmax}\left(\frac{1}{C} \cdot \sum_{c=1}^{C} |F|\right)$$

- Calculate the attention disparity map:

$$D = \left|A(F^T) - A(F^S)\right|$$

- Generate adaptive binary mask:

$$Mask_{i,j} = \begin{cases} 0, \text{if } D_{i,j} < P \\ 1, \text{ Otherwise} \end{cases}$$

- Distillation losses:

$$L_{LD} = \sum_{k}^{C}\sum_{i}^{H}\sum_{j}^{W}\left(F_{i,j,k}^{T} - F_{i,j,k}^{S}\right)^2 i, j\epsilon R_{LD}$$

$$L_{HD} = \sum_{k}^{C}\sum_{i}^{H}\sum_{j}^{W}\left(F_{i,j,k}^{T} - f_{trans}(F_{i,j,k}^{S})\right)^2, i, j\epsilon R_{HD}$$

$$L_{total} = L_{task} + \alpha \cdot L_{HD} + \beta \cdot L_{LD}$$

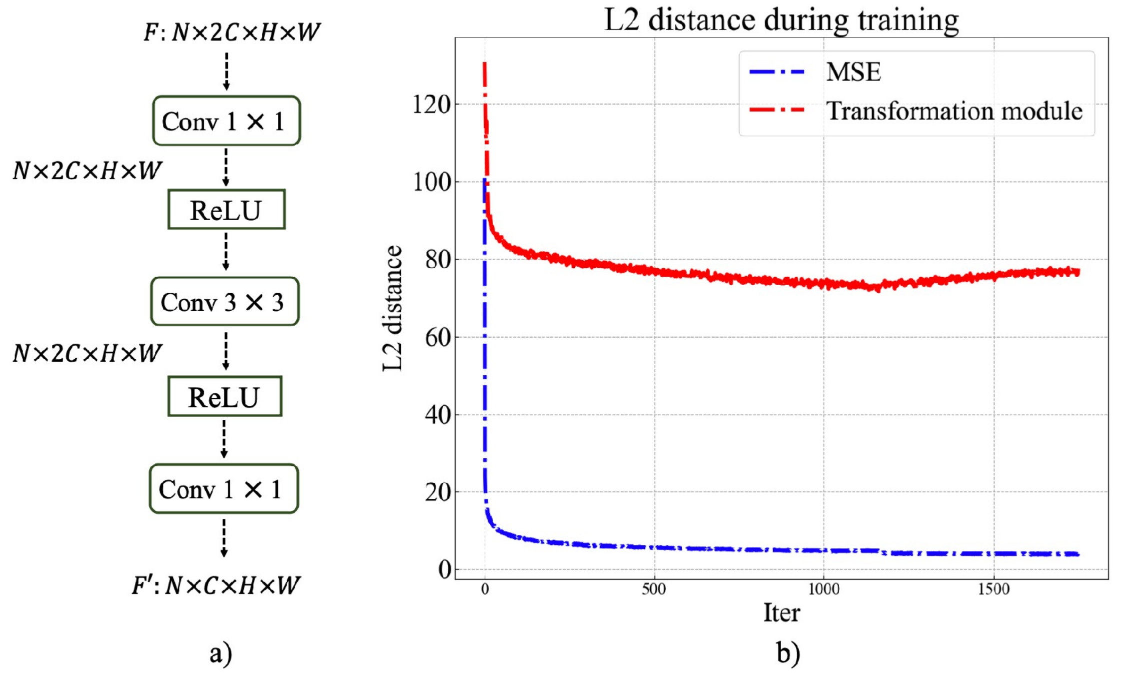


Figure 3. (a). The structure of the transformation module we used. (b). L2 distance during the training with transformation module.

Results and Ablation Study

Table 2.Main results on COCO dataset.

Teacher	Method	schedule	mAP	AP s	AP m	AP L
RetinaNet ResNext101	*RetinaNet ResNet50*	*2x*	*37.4*	*20.0*	*40.7*	*49.7*
	FGD	2x	40.9	23.1	45.1	54.9
	MGD	2x	41.2	23.6	45.3	54.6
	PKD	2x	41.2	23.0	45.4	55.6
	Ours	2x	**41.7(+4.0)**	**24.4**	**46.0**	**55.7**
CascadeMaskRCNN ResNeXt 101	*FaterRCNN ResNet 50*	*2x*	*38.4*	*21.5*	*42.1*	*50.3*
	FGD	2x	42.0	23.8	46.4	55.5
	MGD	2x	42.1	23.7	46.4	56.1
	PKD	2x	41.4	22.7	45.1	56.0
	Ours	2x	**42.4(+4.0)**	**23.9**	**46.8**	**56.3**
Reppoints ResNeXt101	*Reppoints ResNet 50*	*2x*	*38.6*	*22.5*	*42.2*	*50.4*
	FGD	2x	42.0	24.0	45.7	55.6
	MGD	2x	42.3	24.4	46.2	55.9
	PKD	2x	42.4	24.3	46.7	**56.4**
	Ours	2x	**42.7(+4.1)**	**24.8**	**46.7**	56.2

Table 3. Experiments with progressively stronger teacher on COCO dataset.

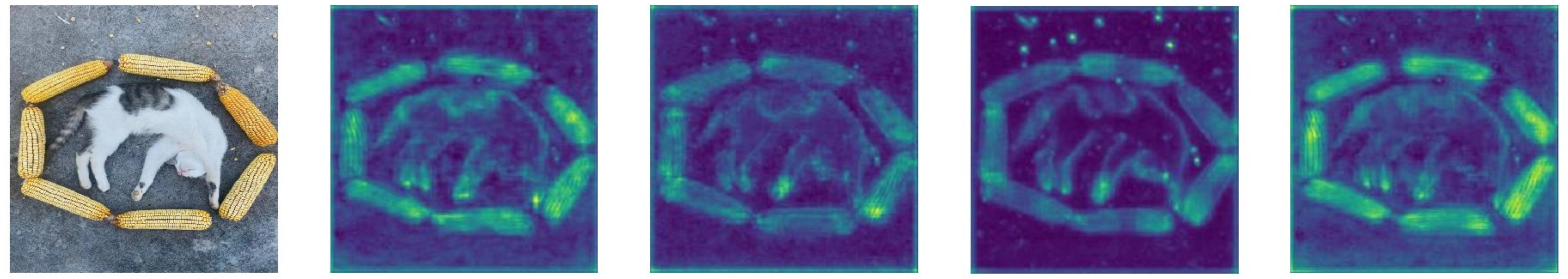
Teacher	Method	Schedule	mAP
FasterRCNN R101 39.8	*student*	*1x*	*37.4*
	PKD	1x	39.6
	Ours	1x	**39.7**
FasterRCNN Rx101 41.2	PKD	1x	40.0
	Ours	1x	**40.5**
MaskRCNN Rx101 42.2	PKD	1x	40.5
	Ours	1x	**41.1**

Table 4. combining DFD with other distillation methods.

Model	Method	Schedule	mAP
RetinaNet RX101 - RetinaNetR50	MGD	1x	40.0
	MGD + Ours	1x	40.3
	PKD	1x	39.9
	PKD + Ours	1x	40.3
	PKD + MGD	1x	39.9

Table 5. Using different constraints in different regions.

HD	LD	Schedule	mAP
MSE	MSE	1x	39.7
TF	TF	1x	40.2
TF	-	1x	39.9
-	TF	1x	39.4
MSE	TF	1x	39.9
TF	TF + MSE	1x	40.2
TF	**MSE**	1x	**40.4**



Original image

Teacher

Student

With MGD

With our method

Figure 4. Visualization of spatial attention after distillation.

Table 6. Experiments of other detectors.

Method	Schedule	mAP
MaskRCNN-SwinS(Teacher)	1x	48.2
MaskRCNN-SwinT(Student)	1x	42.7
PKD	1x	43.9
Ours	1x	**44.4**
Method	Epoch	mAP
YOLOv6-Small(Teacher)	400	44.0
YOLOv6-Tiny(Student)	300	40.6
PKD	300	41.3
Ours	300	**41.7**

Table 7. Results of pose estimation on COCO-Body and segmentation on Cityscapes

Pose estimation	Method	Input Size	mAP
Heatmap Res50	Teacher	256 × 192	71.8
Heatmap MobileNetV2	student	256 × 192	62.0
	CWD	256 × 192	62.2
	Ours	**256 × 192**	**62.6**
Segmentation	Method	Input Size	mAP
PspNet Res101	Teacher	512 × 512	78.34
PspNet Res18	student	512 × 512	69.85
	CWD	512 × 512	73.53
	Ours	**512 × 512**	**73.74**