Distilling Object Detectors with Fine-grained Feature Imitation

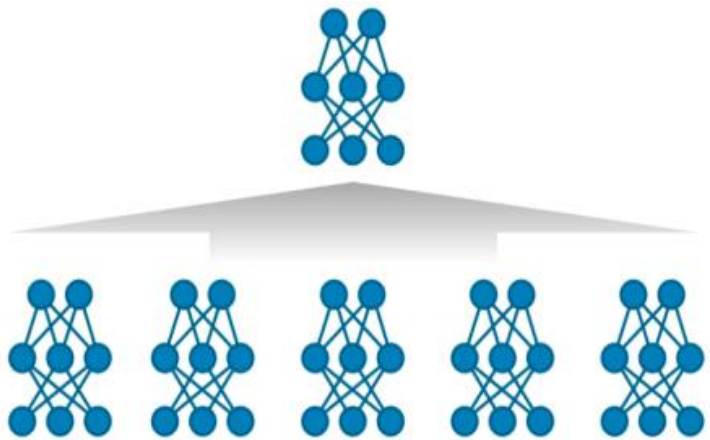
Tao Wang[1] Li Yuan[1] Xiaopeng Zhang[1,2] Jiashi Feng[1]
[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Huawei Noah's Ark Lab, Shanghai, China

CVPR 2019

Wonbeom Jang

Knowledge Distillation

Single Neural Network



Large Ensemble Neural Network

Softmax Output = Knowledge = Soft Label

dog

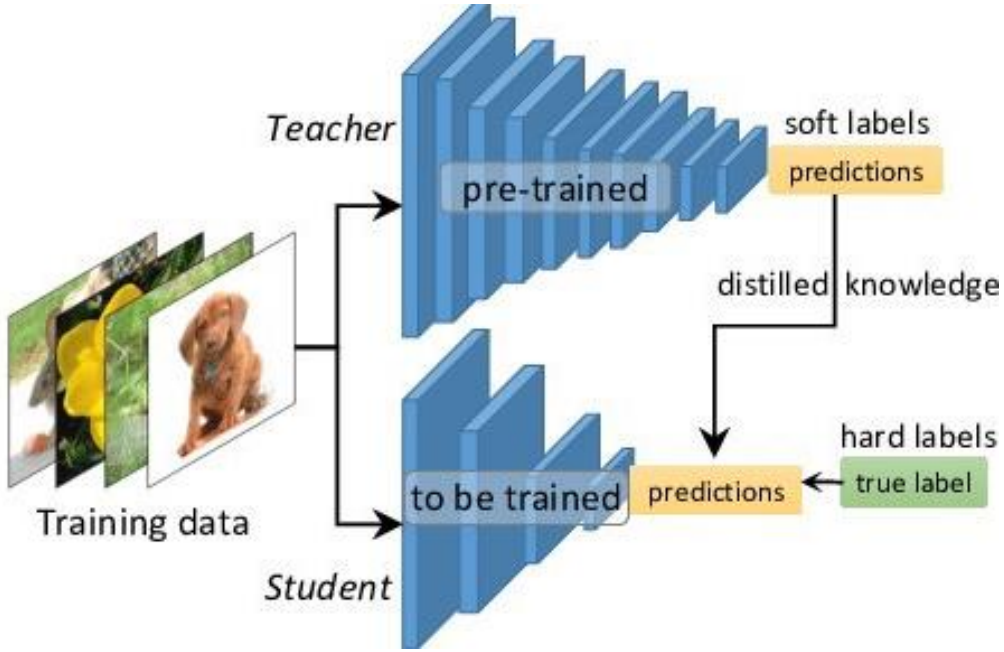
cow	dog	cat	car	
0	1	0	0	original hard targets
10^{-6}	.9	.1	10^{-9}	output of geometric ensemble
.05	.3	.2	.005	softened output of ensemble

https://blog.lunit.io/2018/03/22/distilling-the-knowledge-in-a-neural-network-nips-2014-workshop/

Background

Knowledge Distillation

Architecture



$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

Car	Dog	Cat	Cow
.05	.8	.3	.2

Soft Label

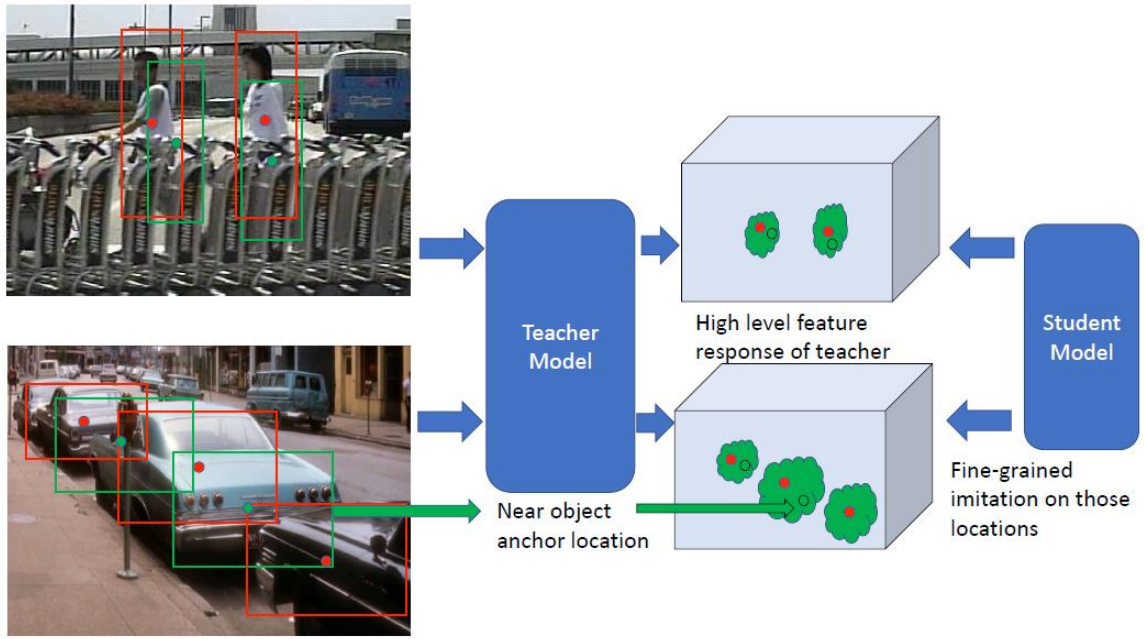
Car	Dog	Cat	Cow
0	1	0	0

Hard Label

Introduction

- Besides, the extreme imbalance of foreground and background instances also makes bounding box annotations less voluminous

- The intuition is that detectors care more about local regions that overlap with ground truth objects while classification models pay more attention to global context.



Introduction

1. We do not rely on softened output of teacher model as in vanilla knowledge distillation of classification model, but depends on a inter-location discrepancy of teacher’s high level feature response.

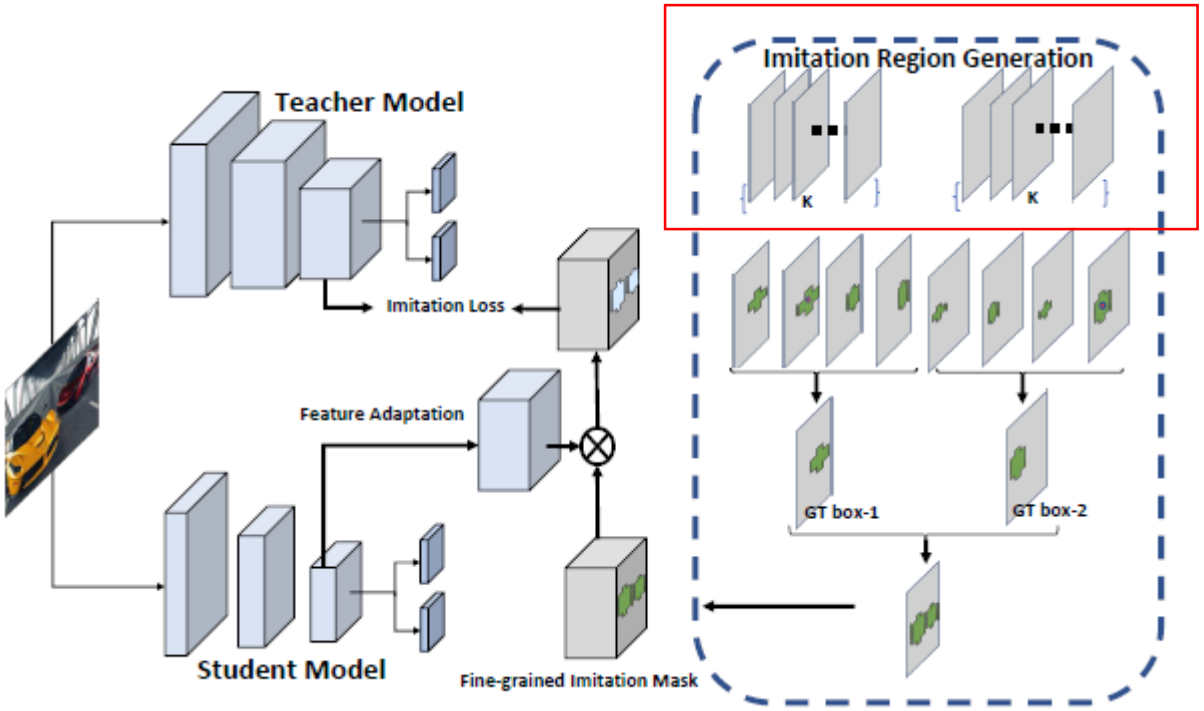
2. Fine-grained feature imitation before classification and localization heads improves both sub-tasks.

3. Our method avoids those noisy less informative background area which leads to degraded performance of full feature imitation, study of the per-channel variance on high level feature maps in Sec 4.4.5 validates this intuition.

Method

Imitation region estimation

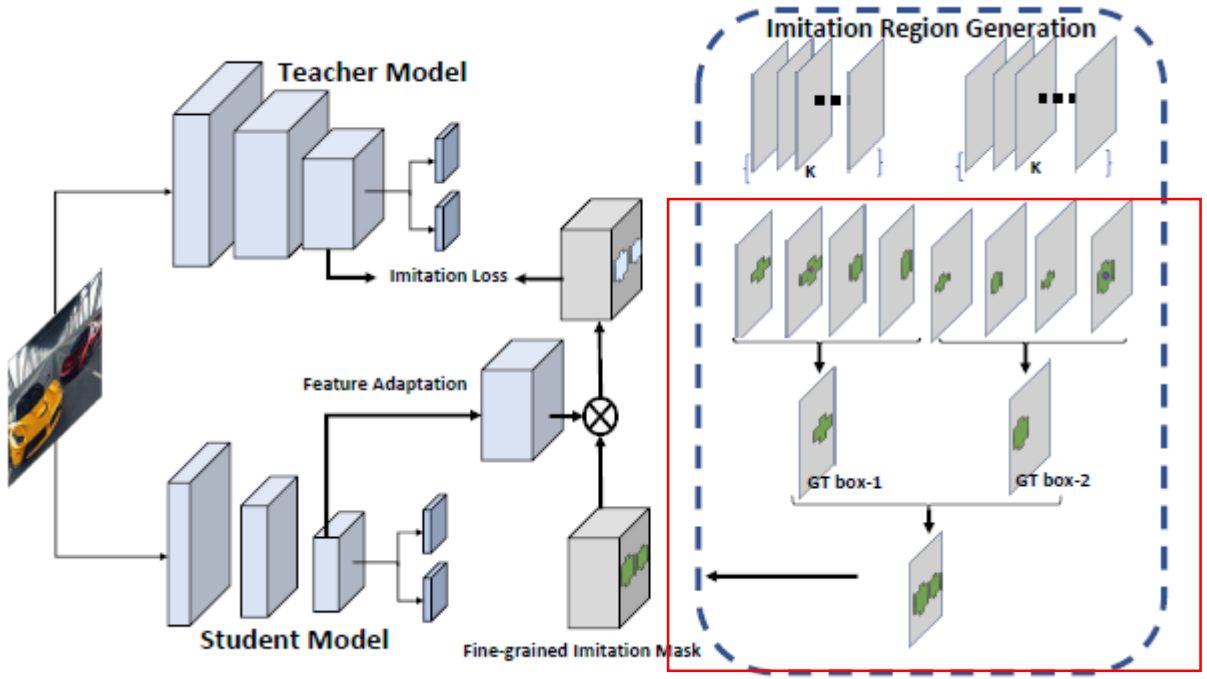
1. Specifically, as shown in Fig. 2, for each ground truth box, we compute the IOU between it and all anchors which forms a $W \times H \times K$ IOU map m. Here W and H denote width and height of the feature map, and K indicates the K preset anchor boxes.



Method

Imitation region estimation

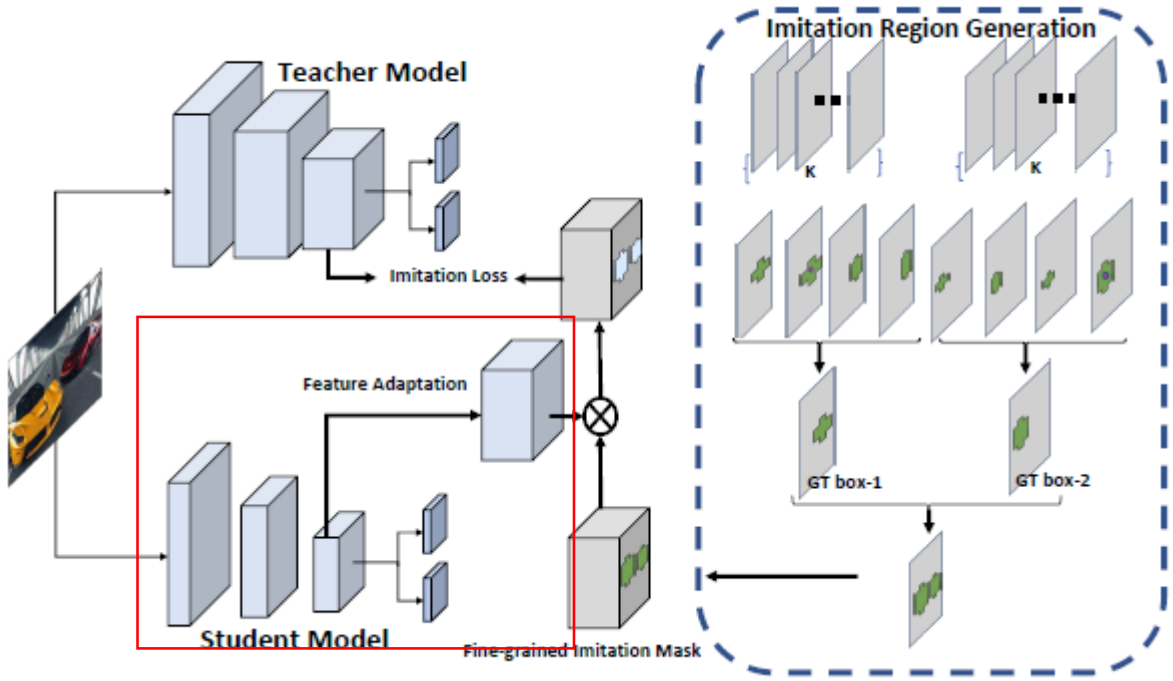
2. Then we find the largest IOU value $M = \max(m)$, times the thresholding factor ψ to obtain a filter threshold $F = \psi \times M$. With F, we filter the IOU map to keep those larger then F locations and combine them with OR operation to get a $W \times H$ mask.



Method

Fine-grained feature imitation

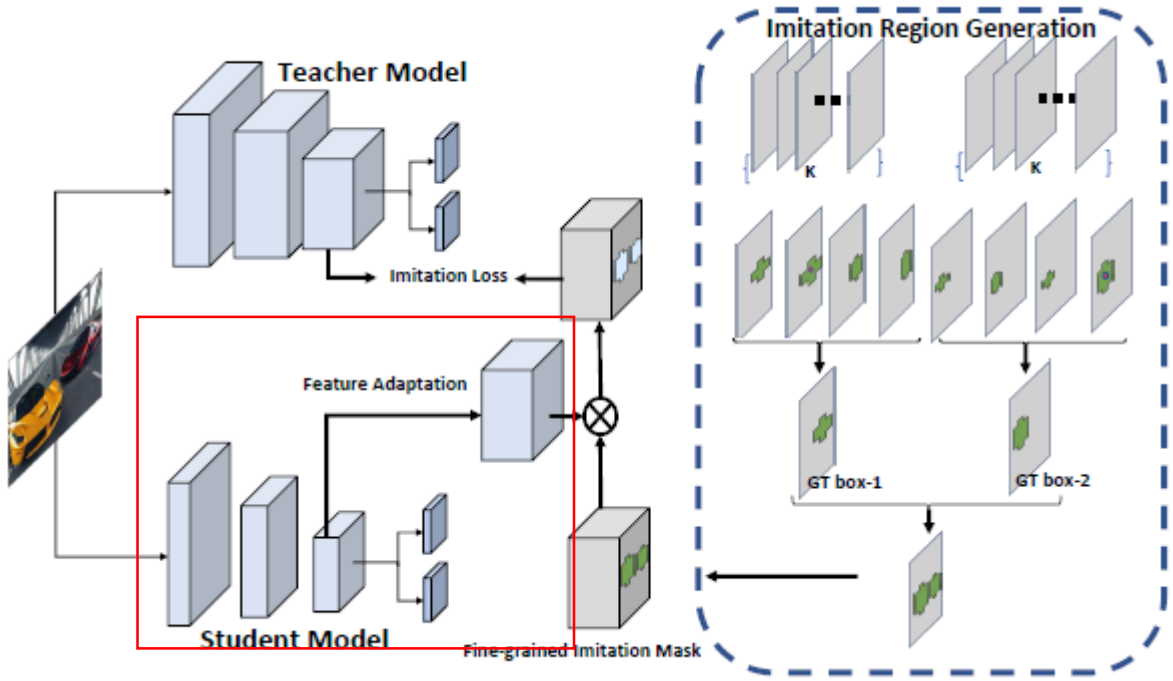
We add a full convolution adaptation layer after corresponding student model before calculating distance metric between student and teacher's feature response, as shown in Figure 2.



Method

Fine-grained feature imitation

We add a full convolution adaptation layer after corresponding student model before calculating distance metric between student and teacher's feature response, as shown in Figure 2.



Method

Fine-grained feature imitation

We add a full convolution adaptation layer after corresponding student model before calculating distance metric between student and teacher's feature response, as shown in Figure 2.

$$L_{imitation} = \frac{1}{2N_p} \sum_{i=1}^{W} \sum_{j=1}^{H} \sum_{c=1}^{C} I_{ij} (f_{\text{adap}}(s)_{ijc} - t_{ijc})^2,$$
$$\text{where } N_p = \sum_{i=1}^{W} \sum_{j=1}^{H} I_{ij}. \tag{2}$$

$$L = L_{gt} + \lambda L_{imitation}, \tag{3}$$

Experiment

Lightweight detector

This detector is based on the Shufflenet which gives excellent classification performance with limited flops and parameters.

(1) We change stride of Conv1 from 2 to 1.

(2) We modify the output channel of Conv1 from 24 to 16, which reduces memory footprint and computation

(3) We reduce the block number of stage-3 from 8 to 6.

(4) We add two additional shufflenet blocks which are trained fromscratch before the regression and classification head.

(5) We employ very simple RPNalike detector which discriminate between classes.

Experiment

Lightweight detector

Models	Flops/G	Params/M	*car*			*pedestrian*			*cyclist*			mAP
			Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	
1×	5.1	1.6	84.56	74.11	65.64	65.28	55.95	50.79	70.39	50.09	46.88	62.63
0.5×	1.5	0.53	76.39	68.35	59.74	63.69	54.34	49.58	64.52	43.67	41.57	57.98
0.5×-I	1.5	0.53	80.56	71.46	61.71	64.18	54.62	49.95	68.25	48.28	45.09	60.46
-	-	-	**+4.2**	**+3.1**	**+2.0**	**+0.5**	**+0.3**	**+0.4**	**+3.7**	**+4.6**	**+3.5**	**+2.5**
0.25×	0.67	0.21	60.36	54.85	46.56	52.41	43.63	39.84	51.35	33.41	31.26	45.96
0.25×-I	0.67	0.21	74.26	61.63	53.94	59.80	50.15	46.28	54.64	38.13	34.84	52.63
-	-	-	**+13.9**	**+6.8**	**+7.4**	**+7.4**	**+6.5**	**+6.4**	**+3.3**	**+4.7**	**+3.6**	**+6.7**
0.25×-F	0.67	0.21	-12.9	-14.5	-11.3	-2.9	-1.9	-1.3	-16.7	-9.3	-9.4	-8.9
0.25×-G	0.67	0.21	+8.8	+2.3	+1.2	+3.1	+0.8	+2.4	-0.5	-0.1	-0.3	+2.0
0.25×-D	0.67	0.21	+3.5	+1.2	+1.3	+1.1	+0.8	+0.3	+0.2	-0.3	-0.1	+0.9
0.25×-ID	0.67	0.21	+10.8	+5.8	+6.3	+6.2	+4.1	+3.6	+2.2	+4.7	+3.1	+5.2

Table 1. Imitation result on the toy detector and results of some comparing methods. 1× is the base detector, 0.5× and 0.25× are directly pruned model trained with ground truth supervision, serving as baselines. -I means with additional proposed imitation loss, -F indicates with full feature imitation, -G means using directly scaled ground truth boxes as imitation region, -D means adding only vanilla distillation loss, -ID indicates the case that both proposed imitation loss and distillation loss are imposed.

Experiment

Imitation with Faster R-CNN

Halved student model

halve channel number of each layer including the fully connected layers to construct the student model

Shallow student network

VGG11 based Faster R-CNN as student and VGG16 based one as teacher; Resnet50 based Faster R-CNN as student and Resnet101 based one as teacher

Multi-layer imitation

we further extend
the experiment to multi-layer imitation with seminal work of Feature Pyramid Networks (FPN)

We compute the imitation region on each layer with corresponding prior anchors, and let student model imitate feature response on each layer

Experiment

Imitation with Faster R-CNN

Model	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
res101	74.4	77.8	78.9	77.5	63.2	62.6	79.2	84.4	85.6	54.5	81.5	68.7	85.7	84.6	77.8	78.6	47.1	76.3	74.9	78.8	71.2
res101h	67.4	73.9	78.6	66.3	52.5	42.4	73.8	80.4	80.1	43.5	71.8	61.9	78.7	81.7	74.4	76.8	42.2	66.9	65.	74.3	62.8
res101h-I	71.2	77.2	80.0	72.9	56.0	50.4	77.1	82.3	85.5	47.4	80.2	59.9	84.3	83.9	73.8	79.1	44.6	70.8	69.4	78.7	70.4
	+3.8	**+3.3**	**+ 1.4**	**+ 6.6**	**+ 3.5**	**+ 8.0**	**+ 3.3**	**+ 1.9**	**+ 5.4**	**+ 3.9**	**+ 8.4**	-2.0	**+ 5.6**	**+ 2.2**	-0.6	**+ 2.3**	**+ 2.4**	**+ 3.9**	**+ 4.4**	**+ 4.4**	**+ 7.6**

Table 2. Imitation with halved student model with Faster R-CNN model on Pascal VOC07 dataset.

Model	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
VGG16	70.4	70.9	78.0	67.8	55.1	53.2	79.6	85.5	83.7	48.7	78.0	63.5	80.2	82.0	74.5	77.2	43.0	73.7	65.8	76.0	72.5
VGG11	59.6	67.3	71.4	56.6	44.3	39.3	68.8	78.4	66.6	37.7	63.2	51.6	58.3	76.4	70.0	71.9	32.2	58.1	57.8	62.9	60.0
VGG11-I	67.6	72.5	73.8	62.8	53.1	49.2	80.5	82.7	76.8	44.8	73.5	64.3	72.6	81.1	75.3	76.3	40.2	66.3	61.8	73.4	70.6
	+8.0	**+5.2**	**+2.4**	**+6.2**	**+8.8**	**+9.9**	**+11.7**	**+4.3**	**+10.2**	**+7.1**	**+10.3**	**+12.7**	**+14.3**	**+4.7**	**+5.3**	**+4.4**	**+8.0**	**+8.2**	**+4.0**	**+10.5**	**+10.6**
res101	74.4	77.8	78.9	77.5	63.2	62.6	79.2	84.4	85.6	54.5	81.5	68.7	85.7	84.6	77.8	78.6	47.1	76.3	74.9	78.8	71.2
res50	69.1	68.9	79.0	67.0	54.1	51.2	78.6	84.5	81.7	49.7	74.0	62.6	77.2	80.	72.5	77.2	40.0	71.7	65.5	75.0	71.0
res50-I	72.0	71.5	80.6	71.1	57.0	52.4	82.1	90.0	82.7	51.6	74.5	66.2	82.3	82.3	75.7	78.3	43.5	79.6	69.1	77.3	72.1
	+2.9	**+2.6**	**+1.6**	**+4.1**	**+2.9**	**+1.2**	**+3.5**	**+5.0**	**+1.0**	**+1.9**	**+0.5**	**+3.6**	**+5.1**	**+2.3**	**+3.2**	**+1.1**	**+3.5**	**+7.9**	**+3.6**	**+2.3**	**+1.1**

Table 3. Imitation with shallow student model on Pascal-VOC07 dataset with Faster R-CNN model.

Experiment

Qualitative performance gain from imitation

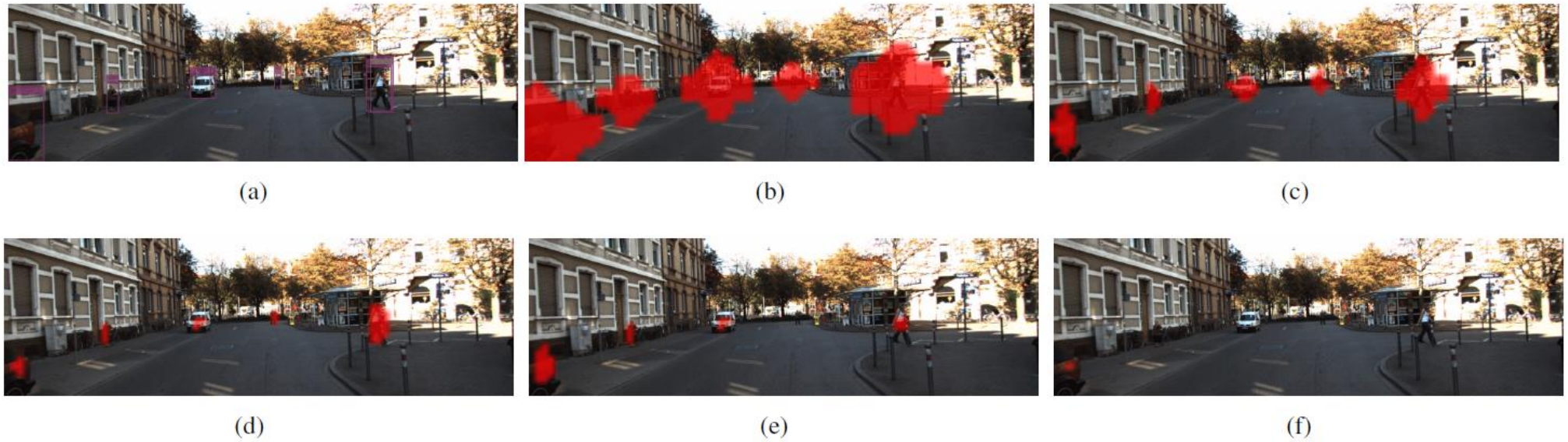


Figure 3. Examples of calculated imitation masks overlaid on input image. Note that the actual masks are calculated on last feature map, we enlarge the mask with corresponding ratio to display on the input image. (a) Original image. (b) $\psi = 0.2$. (c) $\psi = 0.5$. (d) $\psi = 0.8$. (e) Hard-thresh-0.5. (f) Hard-thresh-0.8. Thresh-* indicates different thresholding factor for proposed approach, Hard-thresh-* means using constant threshold of F when filtering the IOU map.

Experiment

Qualitative performance gain from imitation

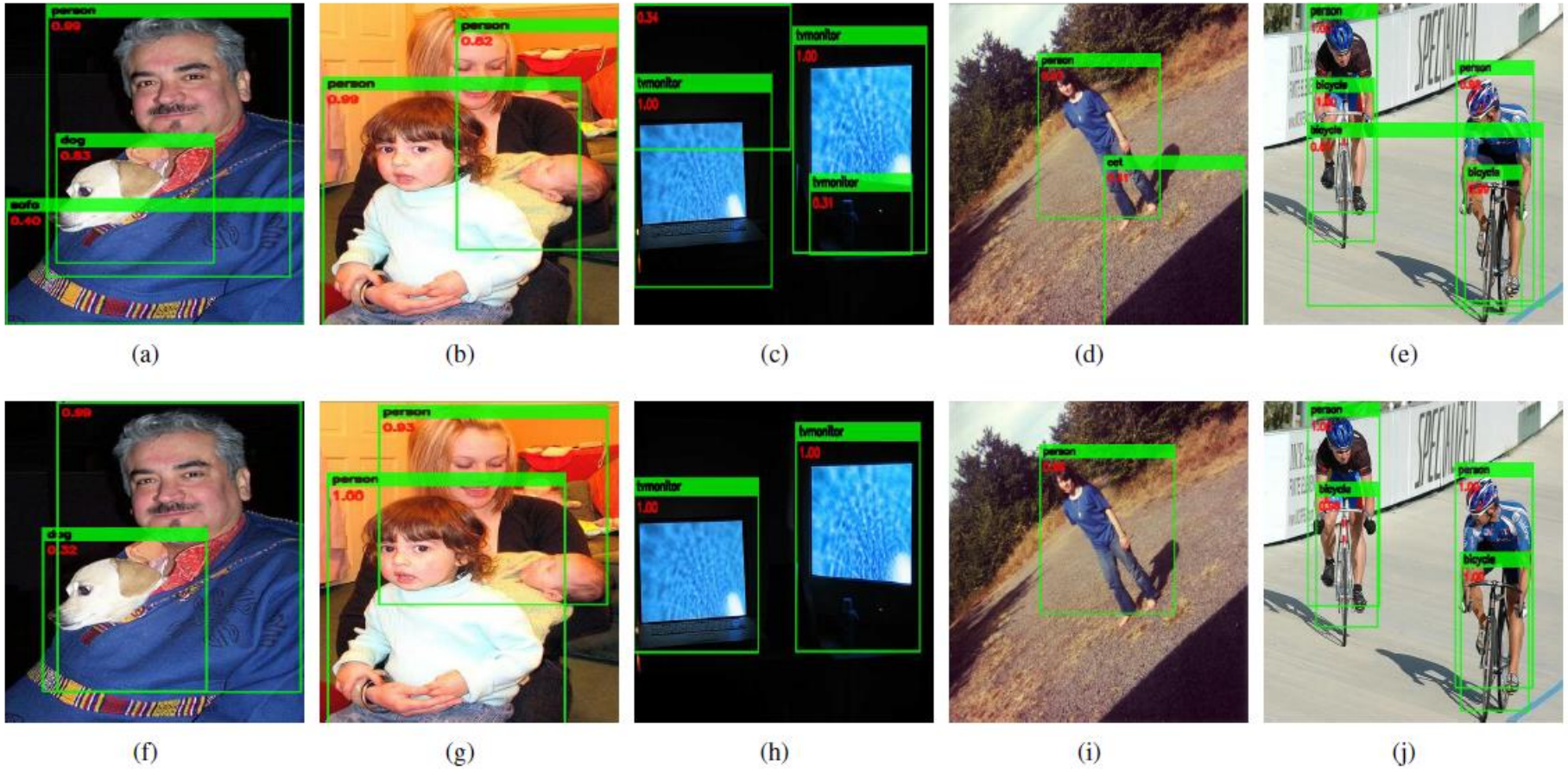
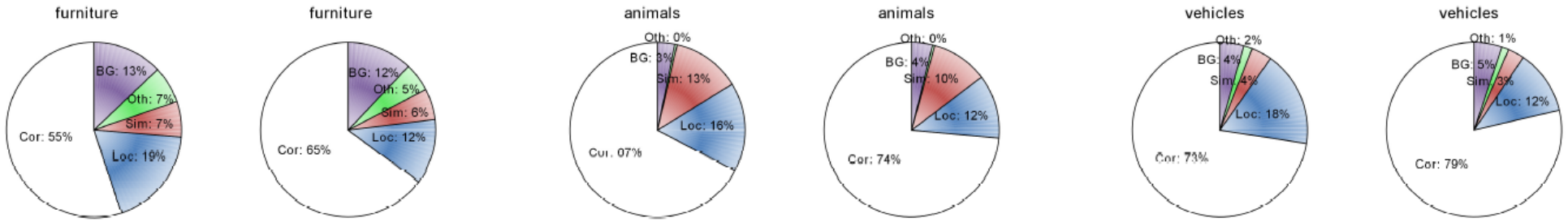


Figure 4. Qualitative results on the gain from imitation learning. The bounding box visualization threshold is set as 0.3. The top row images are student model's output without imitation, the bottom row shows imitated student's output.

Experiment

Qualitative performance gain from imitation



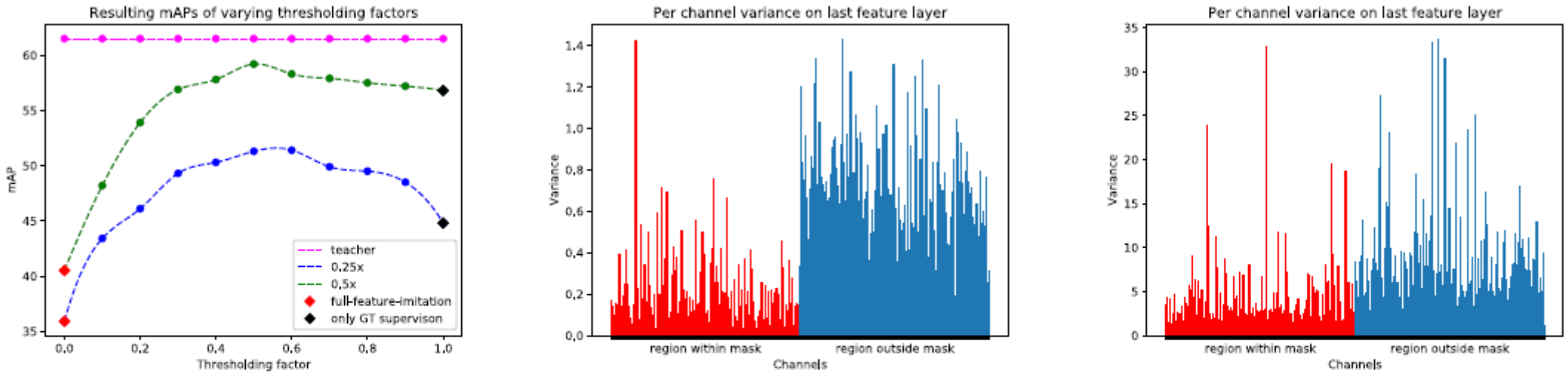
(a) (b) (c)

Figure 5. Imitation gain from error perspective with VGG11 based Faster R-CNN student and VGG16 based teacher on the Pascal VOC07 dataset. For each pair, the left figure corresponds to raw student model, and the right corresponds to imitated student.

1) Stronger localization ability (Loc);
2) less confusion between the same category and other category objects (Sim and Oth);
3) less background induced errors (BG).

Experiment

Quantitative performance gain from imitation

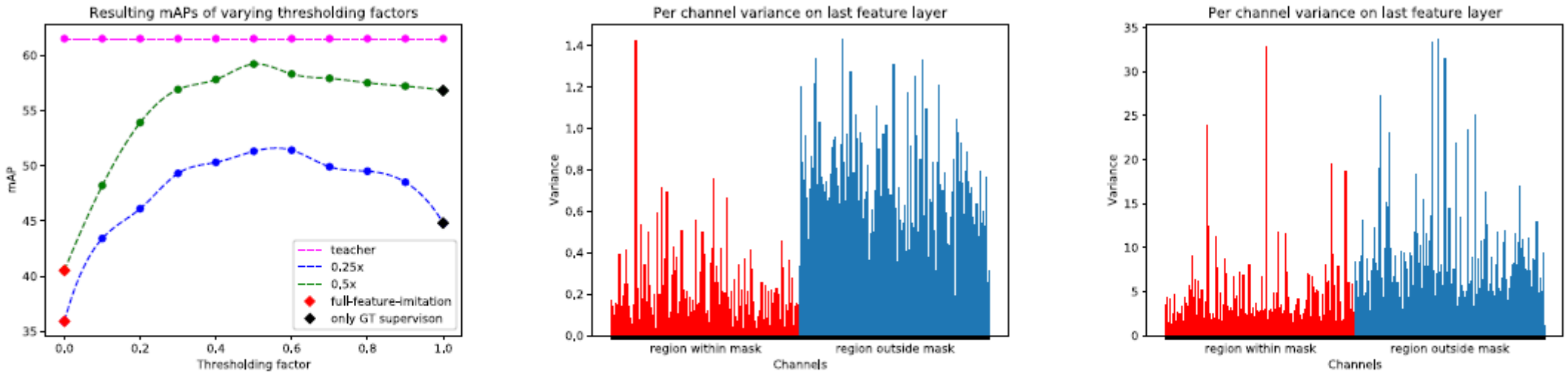


(a) (b) (c)

Figure 6. Results for further investigation of the method. (a) Varying imitation thresholding factor ψ for the toy detector experiment. (b),(c) Per-channel variance on high level feature map of learned teacher model. (b) is calculated with toy detector on KITTI dataset.(c) is calculated with Faster R-CNN on COCO dataset.

Experiment

Quantitative performance gain from imitation



(a) (b) (c)

Figure 6. Results for further investigation of the method. (a) Varying imitation thresholding factor ψ for the toy detector experiment. (b),(c) Per-channel variance on high level feature map of learned teacher model. (b) is calculated with toy detector on KITTI dataset.(c) is calculated with Faster R-CNN on COCO dataset.