SUPPLEMENTARY: WHAT'S WRONG WITH THE ROBUSTNESS OF OBJECT DETECTORS?

Anonymous authors
Paper under double-blind review

The supplementary materials contain two parts:

1. Source code for reproducing the DCM results is in the folder named "code" in the ZIP file. Instructions for running the code are given in the README file in the "code" folder.
2. *This supplementary file* includes related work, more details on DCM and Classification-Ablative Validation, and more clear visualizations zoomed-in of all the results of DCM and Classification-Ablative Validation in the main paper for a better view.

A ATTACK FOR DETECTORS WITH DIFFERENT STRUCTURE

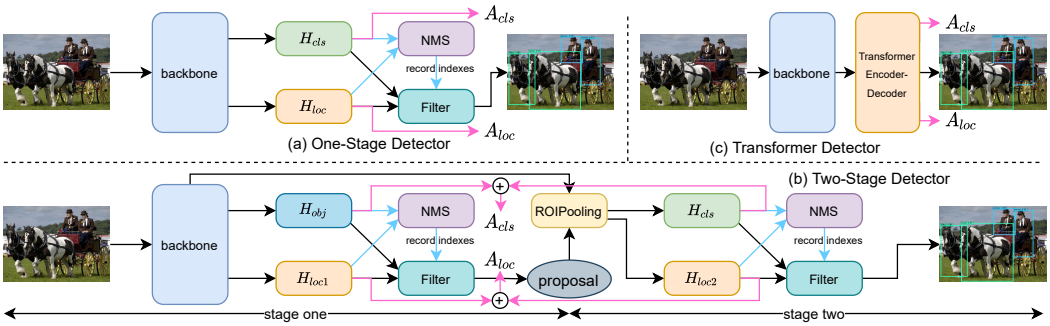

Figure I: Adversarial attack for object detectors with different architectures. Arrows in pink denotes attacks for classification (A_{cls}) or localization (A_{loc}). NMS gives the indexes to be recorded (*record indexes*) based on the predicted bounding boxes and their confidence. $Filter$ is the process to produce the detection bounding boxes based on these indexes, whose confidences are larger than a score threshold.

We have selected four typical object detectors: SSD Liu et al. (2016), Faster-RCNN Ren et al. (2015), YOLOX Ge et al. (2021) and Deformable-DETR Zhu et al. (2021). These detectors can be divided into three types which is One-Stage Detector, Two-Stage Detector and Transformer Detector. The adversarial attack method for these three types of object detectors is shown in Fig. II, and the robustness evaluation in our main paper is based on this standard.

B RELATED WORK

B.1 ADVERSARIAL ATTACK AND DEFENSE ON IMAGE CLASSIFICATION

Deep neural networks have achieved great progress in the classification task. However, these models are demonstrated that they would be completely confused when some imperceptible perturbations were applied to the input Szegedy et al. (2014). Recently, adversarial attack methods are in bloom: gradient-based white box adversarial attack methods (*e.g.*, FGSM Goodfellow et al. (2015) and PDG Madry et al. (2018)), and black box adversarial attack methods (*e.g.*, UPSET Sarkar et al. (2017) and LeBA Yang et al. (2020)). Instead, to resist those adversarial attacks, various defense approaches have been proposed Tramèr et al. (2018); Carlini & Wagner (2017); Liao et al. (2018); Zhang et al. (2020); Qin et al. (2019) and adversarial training becomes prevalent and is widely used to continuously learn adversarial images to neutralize the attack. Despite tremendous progress, few

research works are devoted to the adversarial robustness in the object detection task, especially on adversarial defense. One main difference from the adversarial robustness in image classification is classification models present superior robustness through adversarial training on clean and adversarial images. However, object detectors suffer from a detection robustness bottleneck in adversarial training and are ineffective in balancing robustness on adversarial images and recognition ability on clean images. Thus, in our work, we empirically investigate the adversarial robustness for object detection.

B.2 Adversarial Attack and Defense on Object Detection

With the breakthrough of deep neural networks, object detection has obtained remarkable performance in various scenarios. CNN-based detectors and transformer detectors attract increasing attention, *e.g.*, Faster RCNN Ren et al. (2015), SSD Liu et al. (2016), YOLOX Ge et al. (2021), RetinaNet Lin et al. (2017), and DERT Carion et al. (2020). Even so, they inevitably inherit the vulnerability to attack, with the root in deep neural networks. There are many attack methods that have been proposed specifically for object detectors Xie et al. (2017); Wei et al. (2019); Liu et al. (2019); Chen et al. (2018). In recent years, some works focus on the adversarial robustness of object detectors. MTD Zhang & Wang (2019) as an early attempt regards the adversarial training of object detection as multi-task learning. Classification and localization are both considered to improve the overall robustness of the object detector. Considering that the classes imbalance of the input image will lead to the imbalance of the attack on different categories. CWAT Chen et al. (2021) is proposed to uniformly attack each category in adversarial training to improve the robustness of the detector.

Existing works on the improving robustness of object detectors are suffering from detection robustness bottleneck. Existing works on the robustness of object detectors are suffering from robustness bottlenecks, but their causes and properties are still poorly explored. The main intention of this work is to explore the detection robustness bottleneck and make an attempt to figure out its issues, paving the way for further works.

C Calculation of Detection Confusion Matrix

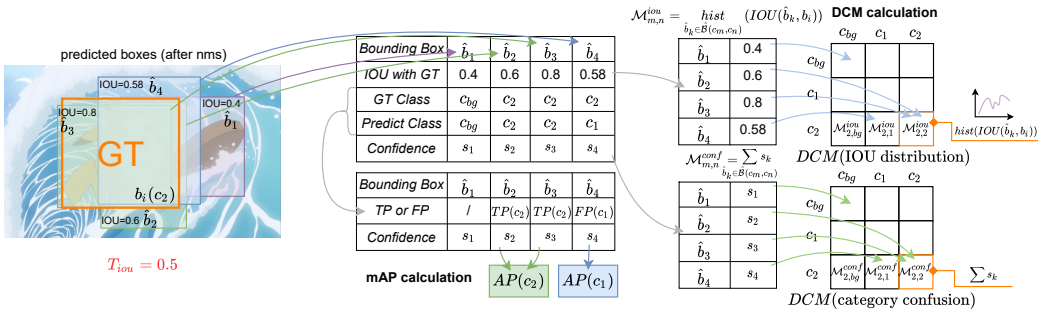

Figure II: More detailed calculation process of Detection Confusion Matrix.

D Evaluation Results under Different Attacks

In our main paper, Fig. 2 provides the performance evaluation of the non-robust and robust models with different detection structures on clean images and adversarial images. In this supplementary, we provide more discussions on the evaluation results in Fig. 2.

Among *four non-robust detection models* (*i.e.*, standard detection models), YOLOX has the highest performance of 83.56% mAP. On adversarial images, all the four detection models show an extremely poor performance: their performance on the A_{cls} adversarial images even degenerates to only lower than 3% mAP! The models perform slightly better on A_{loc} adversarial images than on A_{cls} adversarial images, and even Deformable-DETR achieves an mAP of 13.7%.

Four robust detection models via adversarial training generally have the performance drops by larger than 15%, compared to the non-robust models on clean images. In particular, SSD even loses nearly 30% mAP. Besides, the performance of the robust models on the adversarial images is also limited. None of the four detection models obtain larger than 30% mAP on A_{cls} adversarial images. Although the performance of models on the A_{loc} adversarial images is slightly higher than that on the A_{cls} adversarial images, it also has the performance decline by larger than 40% mAP compared to the non-robust model on clean samples. For example, the standard YOLOX can attain 83.56% mAP on clean images, but the adversarial trained robust YOLOX only has an mAP of 29.10% mAP on A_{cls} adversarial images and 31.90% mAP on A_{loc} adversarial images.

E MORE DETAILS ON IMPLEMENTING DETECTION CONFUSION MATRIX (DCM)

In Sec. 3.1 of the main paper, the details of our proposed DCM have been elaborated. In this supplementary file, we will provide more details on the implementations of reproducing DCM, as shown in Algorithm 1. Specifically, to calculate DCM, it first needs to create two arrays of the size $(C+1) \times (C+1)$ with elements of $list$. To select the bounding boxes for DCM, all the bounding boxes processed by NMS and $filter$ are considered. Those filtered predicted bounding boxes will be matched with the ground-truth (GT) bounding boxes. For a bounding box, we calculate its IOU with all the GT bounding boxes in an image $X[i]$. If the largest IOU is larger than the threshold T_{iou}, we record this IOU value and the index of the matched GT box: b^{iou} and b^{idx}. That is, the GT box at b^{idx} is then matched to this predicted bounding box. For this bounding box, the predicted category and the confidence c^{cls} and c^{conf} are determined based on the score of this predicted bounding box. The category $B[X[i]][b^{idx}].cls$ of the matched GT box is regarded as the true category and c^{cls} as its predicted category. Then DCM can be calculated based on the predicted category and true category of each predicted bounding box.

Algorithm 1: Detection Confusion Matrix

Data: Test images X, GT bounding boxes B, detector backbone f_b, classification header H_{cls}, localization header H_{loc}, IOU matching threshold T_{iou}.

$\mathcal{M}_c \leftarrow list()[C+1, C+1]$; // confidence list matrix.
$\mathcal{M}_b \leftarrow list()[C+1, C+1]$; // bounding box list matrix.
// traversing the test set.
for *i=1* **to** N_X **do**
  $c \leftarrow H_{cls}(f_b(X[i]))$;
  $b \leftarrow H_{loc}(f_b(X[i]))$;
  // get the kept indexes of the boxes.
  $idx \leftarrow NMS(c, b)$;
  $idx \leftarrow filter(idx, c, b)$;
  for $u \in idx$ **do**
    // predicted class and its confidence.
    $c^{cls}, c^{conf} \leftarrow max(c[u])$;
    // the index and IOU of the GT box with the largest IOU of predicted box $b[u]$.
    $b^{idx}, b^{iou} \leftarrow max(IOU(b[u], B[X[i]]))$;
    if $b^{iou} > T_{iou}$ **then**
      $\mathcal{M}_c[B[X[i]][b^{idx}].cls][c^{cls}].append(c^{conf})$;
      $\mathcal{M}_b[B[X[i]][b^{idx}].cls][c^{cls}].append(b^{iou})$;
    end
  end
end
// calculating DCM.
for *i=1* **to** $C+1$ **do**
  for *u=1* **to** $C+1$ **do**
    $\mathcal{M}^{conf}[i][u] \leftarrow sum(\mathcal{M}_c[i][u])$;
    $\mathcal{M}^{iou}[i][u] \leftarrow hist(\mathcal{M}_b[i][u])$;
  end
end

F MORE DETAILS ON IMPLEMENTING CLASSIFICATION-ABLATIVE VALIDATION (CLSAVAL)

The implementation details of our proposed Classification-Ablative Validation are shown in Algorithm 2. An object detector predicts a lot of bounding boxes on an image, and all bounding boxes have their indexes. We perform ClsAVal assuming that bounding boxes at the same indexes predicted by the same structure of the model on the same image with only different perturbations are responsible for predicting the same objects. When performing ClsAVal, we first record the index idx of the model M to be analyzed on the image x and the index idx^R of the reference model M^R on the image $\dot{x}$. In R_{idx} mode, we filter the output of $M(x)$ by idx^R instead of using idx directly for filtering. In the R_{all} mode, we replace the score in the output with the score predicted by $M^R(\dot{x})$ on top of R_{idx}. The score in YOLOX consists of two parts, and thus our method also derives two different modes R_{obj} and R_{conf} to further analyze the problem.

Algorithm 2: Classification-Ablative Validation

Data: Input image x and reference image $\dot{x}$, GT bounding boxes B, detector M, reference detector M^R, IOU matching threshold T_{iou}.

$c, b \leftarrow M(x)$; // prediction of M on x.
// get the kept indexes of the boxes.
$idx \leftarrow NMS(c, b)$;
$idx \leftarrow filter(idx, c, b)$;
$c^R, b^R \leftarrow M^R(\dot{x})$; // prediction of M^R on $\dot{x}$.
// get the kept indexes of the boxes.
$idx^R \leftarrow NMS(c^R, b^R)$;
$idx^R \leftarrow filter(idx^R, c^R, b^R)$;
$\tilde{\mathcal{O}}^R \leftarrow list()$;
if $mode = R_{idx}$ **then**
  for $i \in idx^R$ **do**
    $\tilde{\mathcal{O}}^R.append((c[i], b[i]))$; // Only indexes are from the reference
  end
end
if $mode = R_{all}$ **then**
  for $i \in idx^R$ **do**
    $\tilde{\mathcal{O}}^R.append((c^R[i], b[i]))$; // Both indexes and predicted confidence are from the reference
  end
end
// for YOLOX only.
if $mode = R_{obj}$ **then**
  for $i \in idx^R$ **do**
    $\tilde{\mathcal{O}}^R.append((c^R_{obj}[i] \times c_{conf}[i], b[i]))$;
  end
end
if $mode = R_{conf}$ **then**
  for $i \in idx^R$ **do**
    $\tilde{\mathcal{O}}^R.append((c_{obj}[i] \times c^R_{conf}[i], b[i]))$;
  end
end

G MORE EVALUATION RESULTS FOR EACH CATEGORY

Tab. I∼III provides the experimental results of the four detection models (*i.e.*, SSD, Faster RCNN, YOLOX, and Deformable-DETR) for each category on clean images and adversarial images under two attacks (A_{loc} and A_{cls}). Among them, we conduct experiments of Deformable-DETR on MS-COCO Lin et al. (2014), while the other three models are on PASCAL VOC Everingham et al. (2015).

Table I: The performance of SSD and Faster RCNN in each category on clean images and adversarial images under two attacks (A_{loc} and A_{cls}) on the PASCAL VOC dataset. "STD" indicates the standard detection model which is non-robust. "Robust" denotes the robust detection model obtained via adversarial training.

Method	SSD Liu et al. (2016)						Faster RCNN Ren et al. (2015)					
	Clean		A_{cls}		A_{loc}		Clean		A_{cls}		A_{loc}	
	STD	Robust	STD	Robust	STD	Robust	STD	Robust	STD	Robust	STD	Robust
aeroplane	81.4	57.3	0.4	49.1	2.3	46.4	67.9	56.6	0.0	24.8	0.3	32.1
bicycle	85.7	66.7	0.9	34.8	6.0	44.2	77.5	64.6	0.0	14.7	3.0	30.2
bird	75.3	36.7	6.1	23.6	0.9	21.6	67.1	41.6	0.0	4.7	0.1	11.0
boat	69.4	28.7	0.1	14.8	0.6	17.3	54.9	35.5	0.2	9.7	1.5	11.7
bottle	50.2	19.7	0.8	10.6	9.2	10.3	53.4	34.7	7.3	9.1	9.1	7.1
bus	83.7	61.5	1.1	51.3	11.4	48.4	75.4	64.5	0.0	15.9	0.6	29.3
car	85.4	71.2	0.9	47.6	11.7	54.9	83.5	72.2	0.0	28.4	1.8	42.6
cat	87.5	48.1	0.1	36.0	9.3	23.4	85.2	58.9	0.0	4.2	1.5	10.0
chair	61.6	31.4	1.0	16.8	9.1	17.6	49.4	38.5	0.0	1.0	0.1	7.0
cow	83.0	37.9	0.1	11.9	3.8	15.5	77.2	54.8	0.0	0.3	0.6	13.6
diningtable	79.4	47.7	1.5	38.8	4.0	35.1	60.1	55.5	0.8	12.7	9.4	26.3
dog	84.4	49.3	0.3	29.1	9.1	31.8	80.5	53.3	0.0	3.6	0.2	13.3
horse	86.1	66.8	0.4	37.9	1.4	44.9	82.7	69.0	0.1	8.2	0.5	28.0
motorbike	84.2	62.4	0.9	28.9	1.5	44.2	74.9	62.6	0.0	14.9	1.0	33.3
person	78.0	57.7	5.0	41.1	9.6	42.5	77.7	65.8	0.0	15.4	1.0	23.4
pottedplant	49.3	20.8	0.1	4.0	3.1	9.7	41.4	30.8	0.0	0.4	0.4	9.4
sheep	75.5	32.5	0.1	10.9	1.9	18.3	69.8	53.2	0.0	0.6	2.6	12.7
sofa	78.9	58.8	0.3	51.6	2.3	43.6	64.4	48.2	0.0	11.5	0.1	16.8
train	85.6	62.7	1.3	39.0	1.5	39.7	73.8	53.8	0.0	12.2	0.3	19.8
tvmonitor	75.3	50.4	9.1	42.3	6.1	37.8	73.7	58.9	0.3	19.0	0.7	24.0

Table II: The performance of YOLOX in each category on clean images and adversarial images under two attacks (A_{loc} and A_{cls}) on PASCAL VOC dataset. "STD" indicates the standard detection model which is non-robust. "Robust" denotes the robust detection model obtained via adversarial training.

Method	YOLOX Ge et al. (2021)					
	Clean		A_{cls}		A_{loc}	
	STD	Robust	STD	Robust	STD	Robust
aeroplane	89.4	67.0	9.1	22.4	3.3	24.5
bicycle	89.6	74.5	9.1	31.9	9.7	46.3
bird	82.3	53.3	9.1	2.8	1.5	16.7
boat	77.1	51.8	0.1	5.9	1.6	12.8
bottle	79.6	60.2	4.6	10.3	9.2	27.1
bus	88.4	70.0	0.9	32.7	11.0	51.0
car	89.9	82.4	9.9	44.1	12.4	60.3
cat	85.9	64.9	0.2	17.1	1.2	34.9
chair	72.3	48.5	0.1	7.7	1.2	25.8
cow	87.8	68.7	0.3	4.2	9.3	30.7
diningtable	79.1	65.5	0.4	21.4	3.3	34.9
dog	85.4	59.0	0.3	8.0	4.8	26.9
horse	88.4	76.6	1.0	26.8	4.8	43.9
motorbike	89.5	69.5	0.7	31.8	9.8	45.1
person	88.0	79.0	1.0	33.0	5.0	48.3
pottedplant	66.1	40.0	0.1	4.0	4.6	14.2
sheep	82.6	62.7	0.1	6.0	2.4	28.3
sofa	82.9	57.1	0.1	6.4	0.7	33.9
train	85.8	68.9	0.7	22.1	9.6	30.7
tvmonitor	84.5	63.7	9.2	23.0	5.4	44.5

Table III: The performance of Deformable-DETR in each category on clean images and adversarial images under two attacks (A_{loc} and A_{cls}) on the MS-COCO dataset. "STD" indicates the standard detection model which is non-robust. "Robust" denotes the robust detection model obtained via adversarial training.

Method	Deformable-DETR Zhu et al. (2021)												
	Clean		A_{cls}		A_{loc}			Clean		A_{cls}		A_{loc}	
	STD	Robust	STD	Robust	STD	Robust		STD	Robust	STD	Robust	STD	Robust
person	78.5	59.7	9.3	29.5	26.9	38.3	wine glass	55.1	29.1	0.5	7.8	13.1	13.2
bicycle	56.4	31.2	0.7	5.9	9.2	14.3	cup	56.7	30.4	0.2	4.4	10.8	14.8
car	63.2	43.3	1.4	16.0	12.3	23.9	fork	51.8	16.7	0.0	3.1	6.6	6.7
motorcycle	72.5	46.5	0.7	11.2	15.3	27.8	knife	27.6	9.6	0.0	0.6	2.0	3.1
airplane	83.7	63.4	4.2	26.2	31.4	40.2	spoon	27.1	6.8	0.0	0.2	0.9	2.3
bus	81.1	63.2	2.0	22.3	33.2	46.2	bowl	54.9	24.9	0.2	4.4	13.2	12.2
train	83.0	59.8	7.5	23.4	35.8	42.9	banana	42.6	25.5	0.2	4.0	9.2	13.9
truck	55.4	20.5	0.4	3.1	10.4	9.1	apple	28.8	16.5	0.0	1.6	1.5	7.1
boat	49.9	23.7	0.1	2.8	2.7	5.5	sandwich	50.4	21.4	0.1	1.8	7.5	12.1
traffic light	50.2	29.1	0.6	6.2	2.4	10.1	orange	40.0	28.0	0.1	8.3	10.8	20.6
fire hydrant	84.2	68.1	3.1	29.8	24.7	52.6	broccoli	41.9	29.9	0.2	5.1	6.1	14.1
stop sign	74.4	61.4	4.4	35.0	41.6	55.2	carrot	34.2	20.1	0.1	2.0	3.2	7.7
parking meter	60.2	41.0	0.0	5.8	15.9	17.5	hot dog	51.5	21.7	0.0	3.7	10.6	12.0
bench	37.8	15.7	0.3	2.3	5.8	7.4	pizza	70.7	47.8	1.5	18.7	28.8	38.4
bird	56.0	30.9	0.0	3.6	6.4	16.4	donut	63.0	28.7	0.1	2.5	11.2	12.6
cat	90.7	52.7	0.7	8.4	22.6	27.6	cake	59.6	28.0	0.2	3.3	11.0	15.0
dog	83.0	47.6	0.4	4.6	18.4	32.9	chair	45.3	17.8	0.0	1.4	4.6	6.5
horse	83.4	48.1	2.1	12.3	20.6	30.7	couch	56.3	36.9	0.6	7.6	17.3	22.5
sheep	79.7	42.1	0.5	7.1	15.8	25.4	potted plant	46.6	20.6	0.1	3.8	4.5	8.3
cow	80.9	47.5	0.6	11.2	19.6	26.2	bed	60.5	33.0	1.4	7.5	24.9	22.5
elephant	88.3	52.0	1.8	4.1	28.9	31.6	dining table	37.3	22.1	1.5	13.3	17.0	19.8
bear	89.1	52.1	0.4	5.5	34.9	36.6	toilet	78.2	51.3	0.4	17.8	23.5	34.2
zebra	91.8	77.9	18.3	42.1	43.5	60.5	tv	76.4	50.2	0.6	10.7	24.0	32.9
giraffe	89.7	72.9	17.6	37.0	45.7	54.7	laptop	76.0	45.6	0.8	12.4	23.7	26.5
backpack	28.0	10.2	0.0	1.5	0.8	3.2	mouse	76.9	52.9	0.4	5.4	18.7	21.4
umbrella	61.9	35.6	0.3	6.8	10.1	20.0	remote	46.1	14.8	0.1	1.0	3.8	3.3
handbag	26.5	7.2	0.1	1.3	1.5	2.0	keyboard	71.6	52.9	1.6	15.7	19.0	33.8
tie	52.4	28.5	1.1	7.4	5.6	12.4	cell phone	49.7	24.3	0.1	1.7	5.7	10.9
suitcase	63.2	20.8	0.0	1.5	7.4	8.4	microwave	71.9	42.2	3.3	8.6	14.8	20.1
frisbee	88.1	70.1	0.4	17.7	21.6	38.7	oven	55.5	21.8	1.9	4.1	15.4	11.6
skis	46.3	19.9	0.3	7.8	1.3	7.4	toaster	38.0	15.8	0.0	0.0	4.1	2.4
snowboard	48.6	20.2	0.2	0.7	2.3	5.1	sink	58.6	31.4	0.4	3.4	8.0	13.4
sports ball	59.6	30.7	0.7	13.5	7.1	19.1	refrigerator	70.0	42.6	1.3	12.0	22.3	26.6
kite	63.8	42.8	0.6	12.8	8.8	29.4	book	23.8	8.7	0.0	0.7	1.2	2.7
baseball bat	60.2	14.9	0.4	2.6	2.9	3.7	clock	73.5	58.1	1.2	28.8	13.5	38.8
baseball glove	60.2	36.1	0.2	8.4	4.1	17.5	vase	56.7	29.9	0.5	3.3	9.1	12.9
skateboard	75.2	46.0	1.9	11.0	9.7	21.4	scissors	35.4	12.4	0.2	3.0	10.4	5.2
surfboard	60.5	32.2	0.3	6.2	7.1	12.4	teddy bear	71.1	34.6	0.1	4.8	18.0	19.8
tennis racket	76.2	54.2	2.3	23.2	15.4	31.8	hair drier	15.3	18.8	0.0	0.0	2.6	0.0
bottle	54.1	28.1	0.4	2.6	7.7	10.3	toothbrush	39.6	10.6	0.1	4.7	1.4	4.4

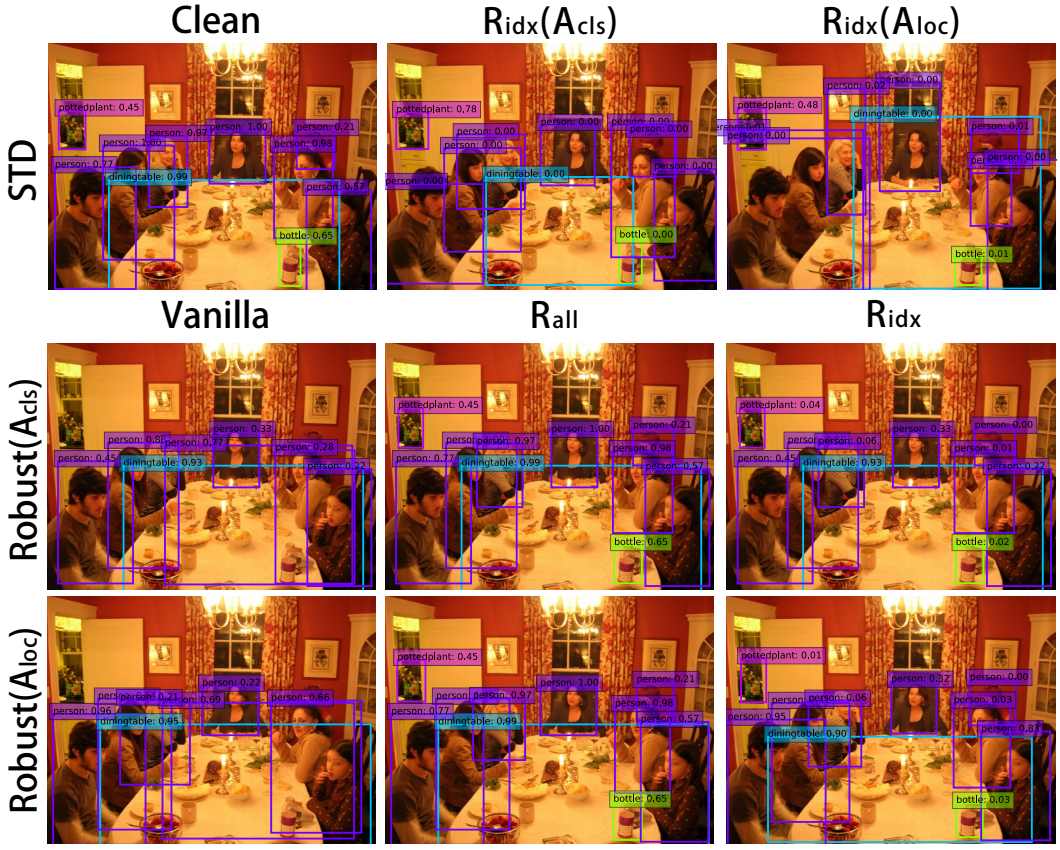

Figure III: Classification-Ablative Validation for standard and robust SSD models.

H VISUALIZATION OF DCM AND CLSAVAL

The results of DCM and ClsAVal have been shown in Fig. 6∼13 in the main paper. In this supplementary file, we provide corresponding visualizations (in Fig. III∼XIV) again for a better view (including the ClsAVal of RPN in standard and robust Faster R-CNN), due to so many object categories, with the same results as those in the main paper. In the main paper, SSD, Faster-RCNN and YOLOX are evaluated on the PASCAL VOC dataset, while Deformable-DETR is on the MS-COCO dataset. Thus, for Deformable-DETR, we select objects in MS-COCO whose category is the same as the PASCAL VOC dataset to calculate DCMs in the main paper, as shown in Fig.13, Fig. XIII and Fig. XIV. In addition, we also provide the visualizations of DCMs on all the 80 object categories in MS-COCO for Deformable-DETR, as shown in Fig. XV∼Fig. XXVIII. Fig. XV∼Fig. XXVIII present the results of DCM and ClsAVal for standard Deformable-DETR and robust Deformable-DETR on clean images and adversarial images under two attacks (A_{loc} and A_{cls}), respectively. The detailed analysis based on these experimental results is presented in the Sec. 3 and Sec. 4 of main paper.

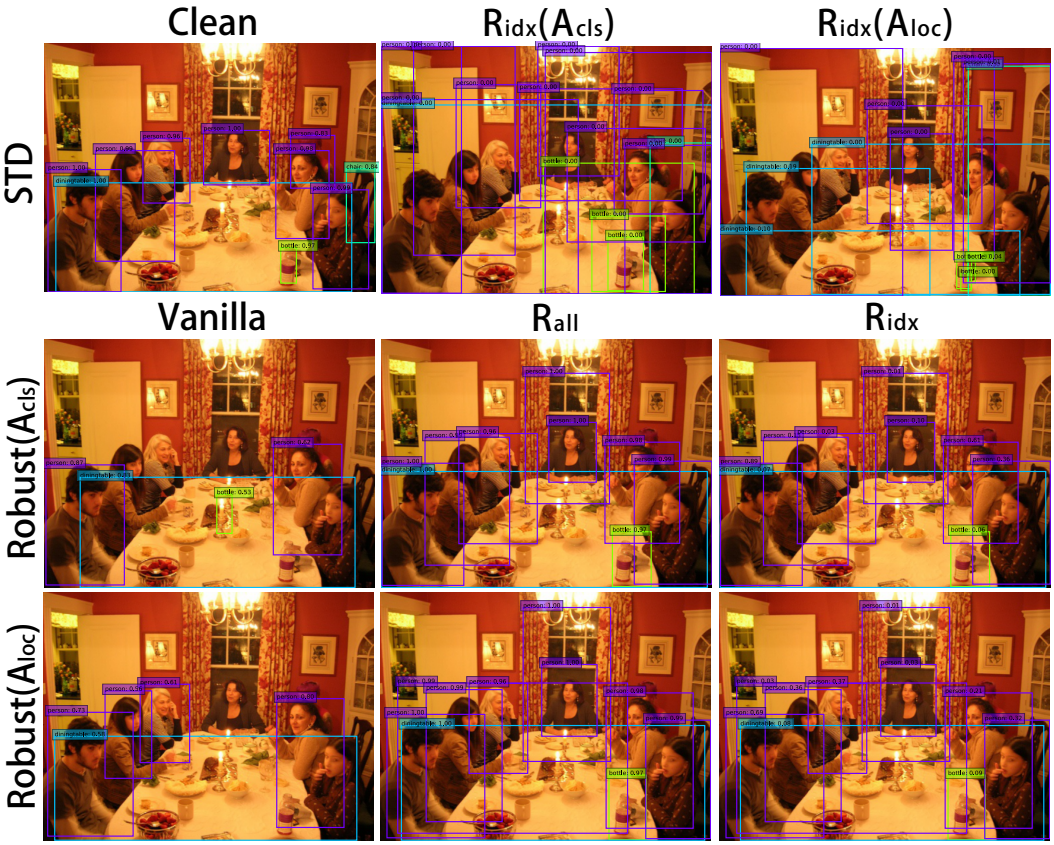

Figure IV: Classification-Ablative Validation for standard and robust Faster R-CNN models.

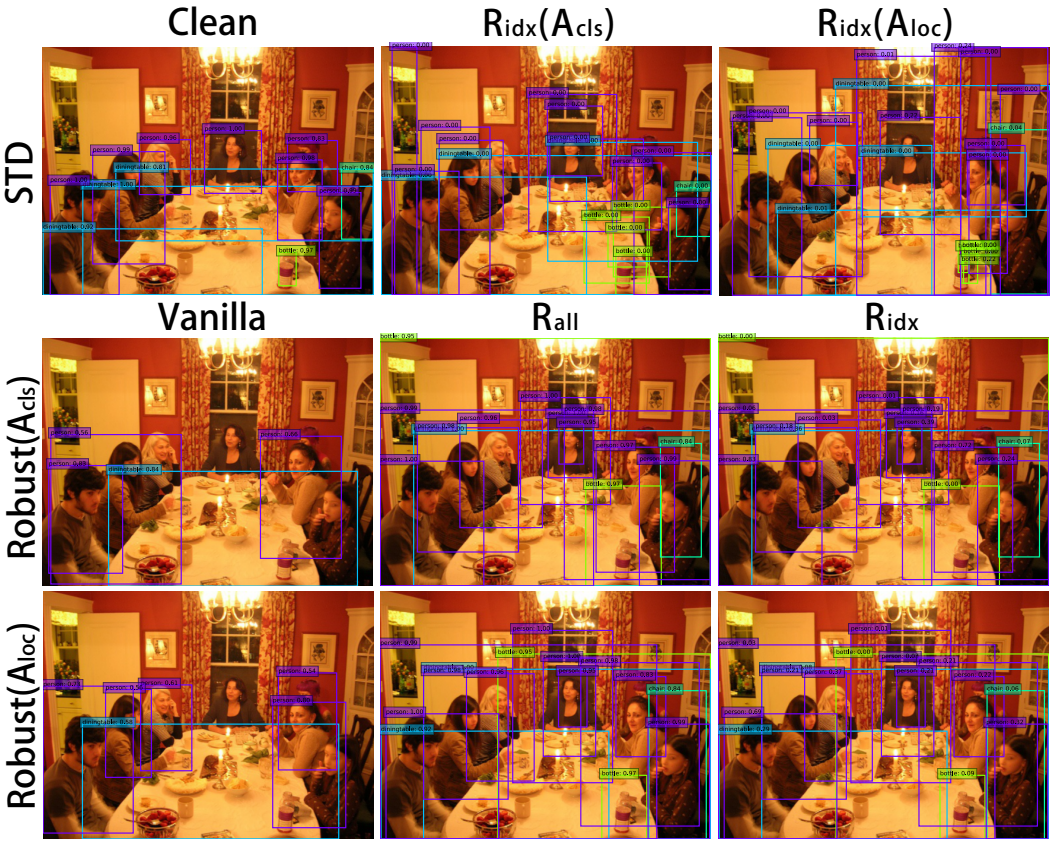

Figure V: Classification-Ablative Validation for RPN in standard and robust Faster R-CNN.

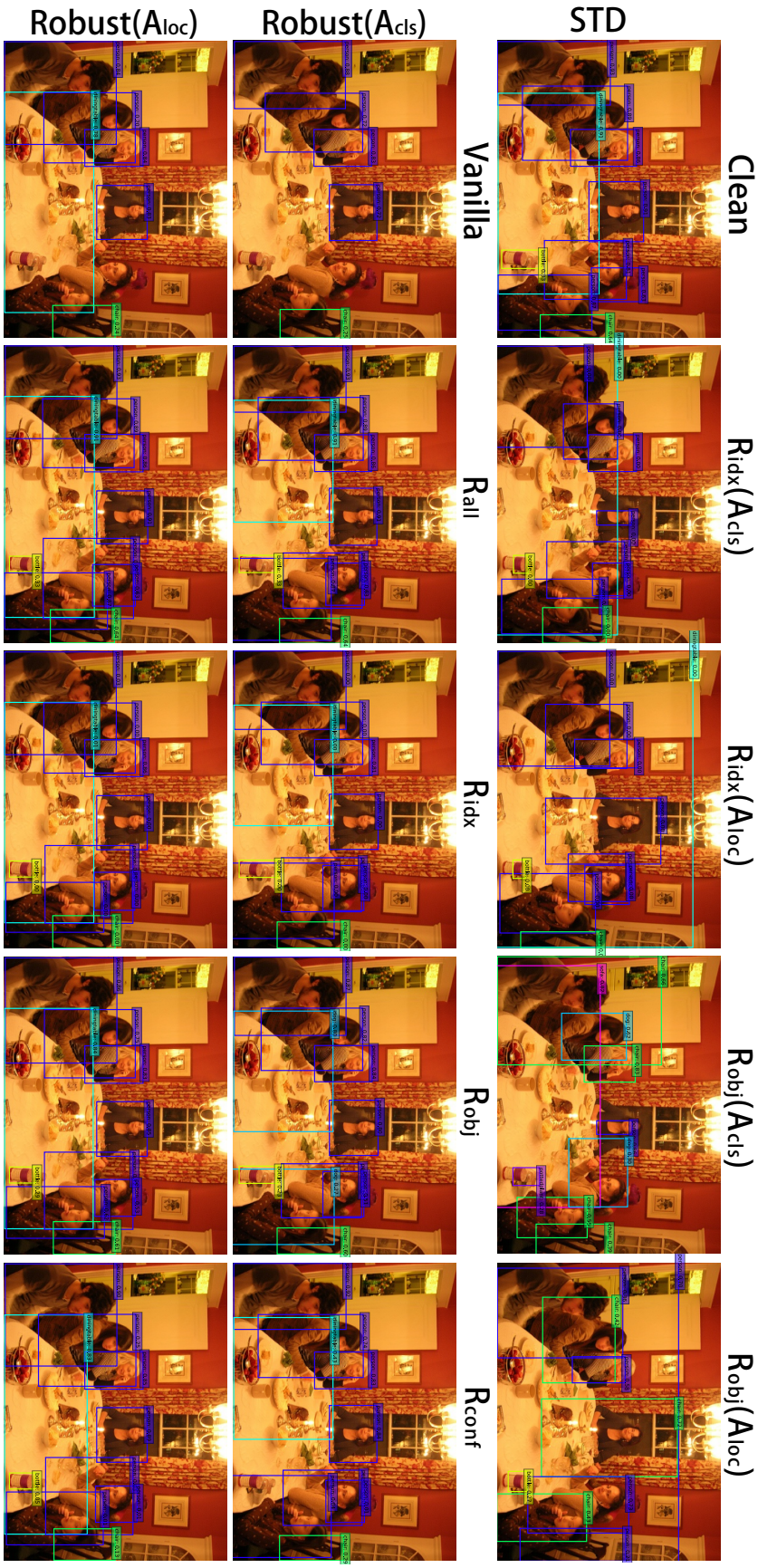

Figure VI: Classification-Ablative Validation for standard and robust YOLOX models.

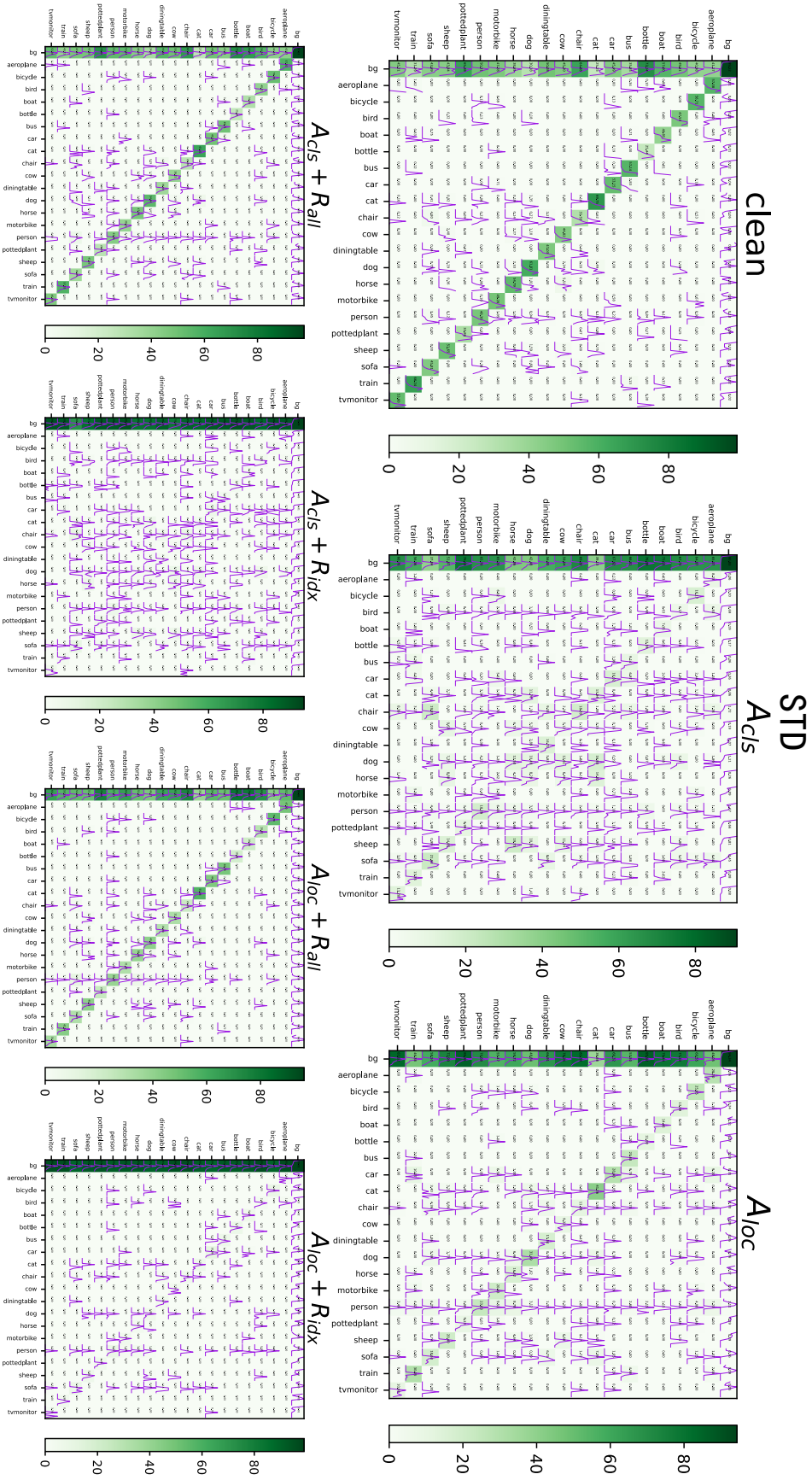

Figure VII: DCM of the standard SSD Detector.

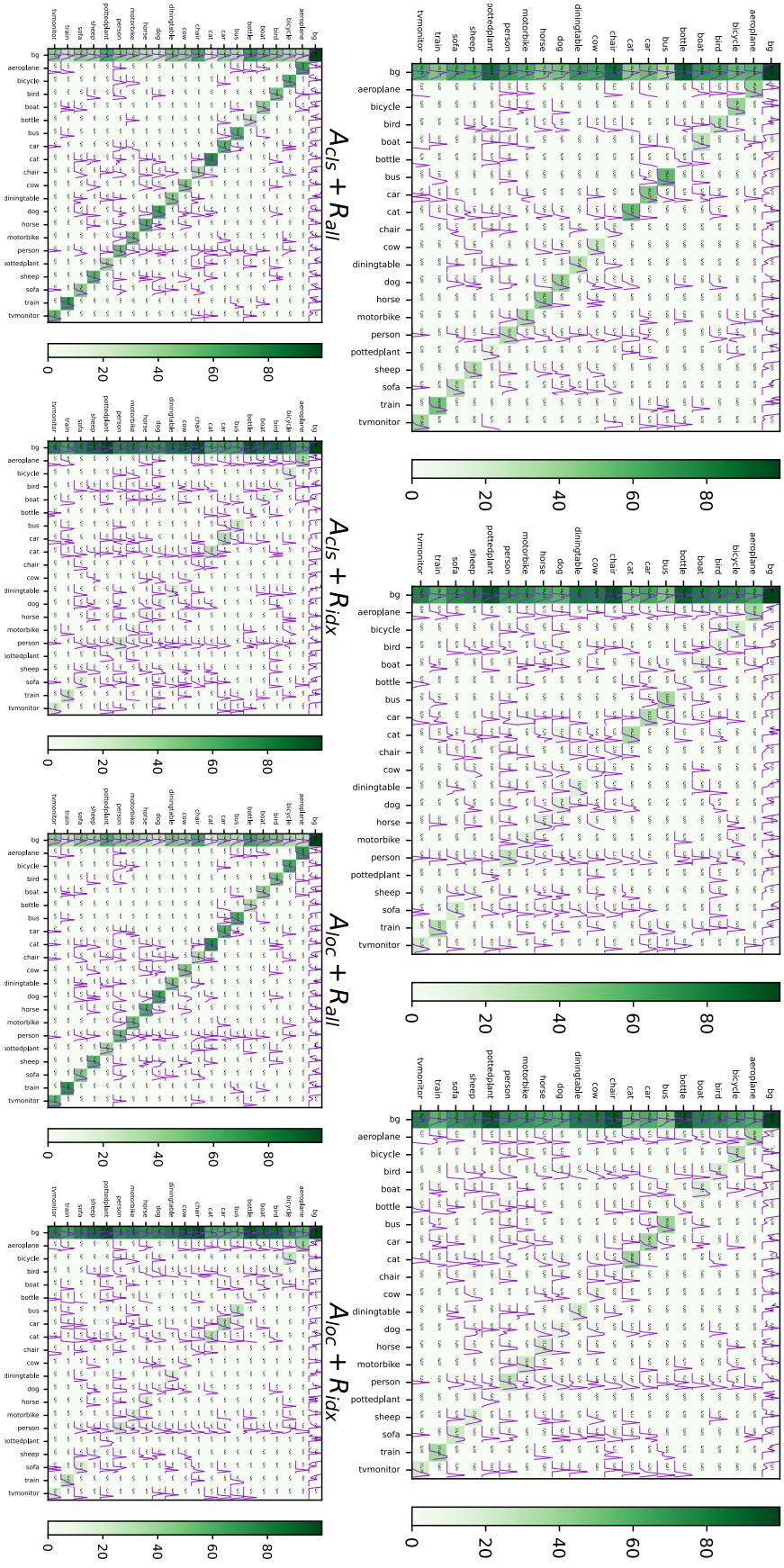

Figure VIII: DCM of the robust SSD Detector.

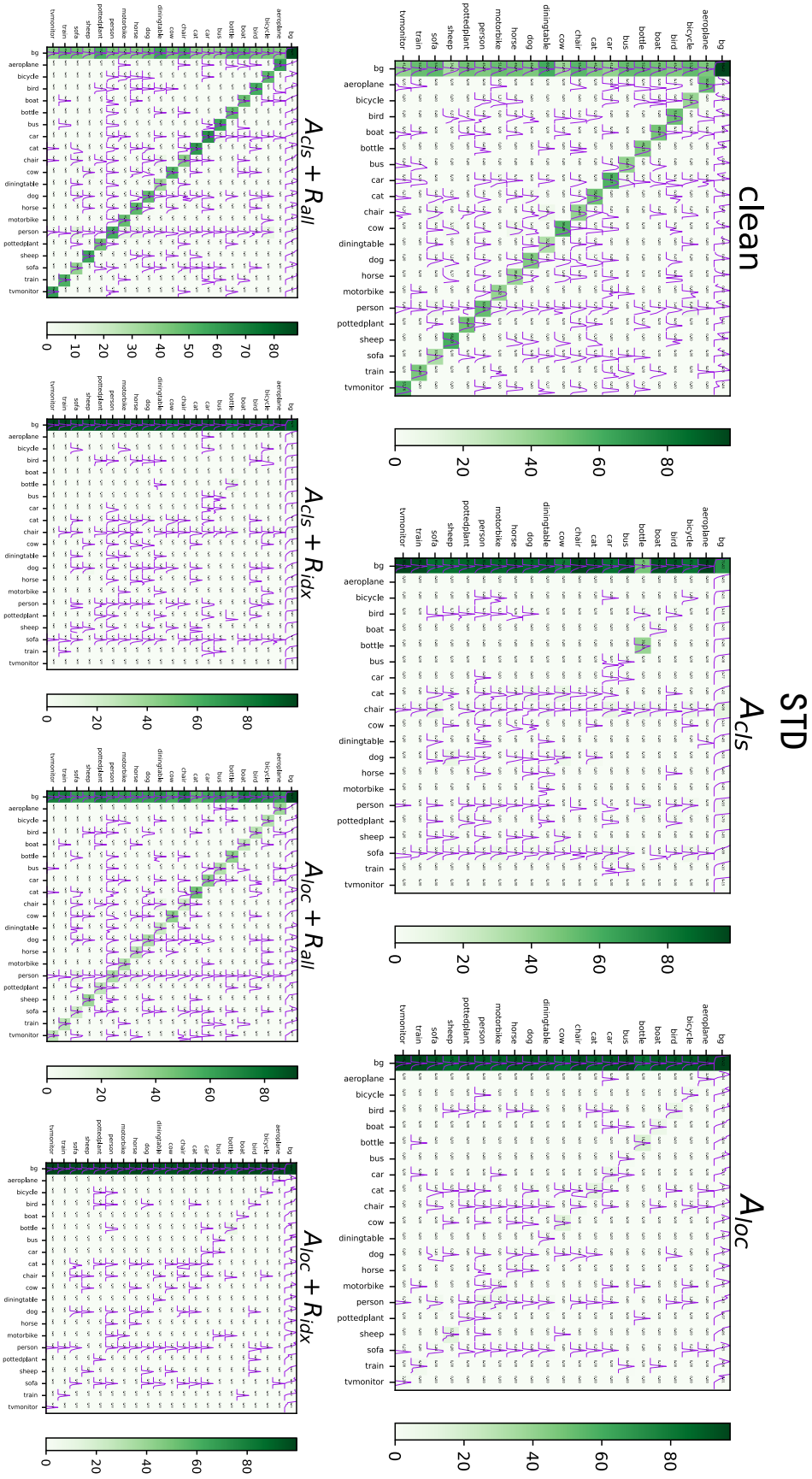

Figure IX: DCM of the standard Faster RCNN Detector.

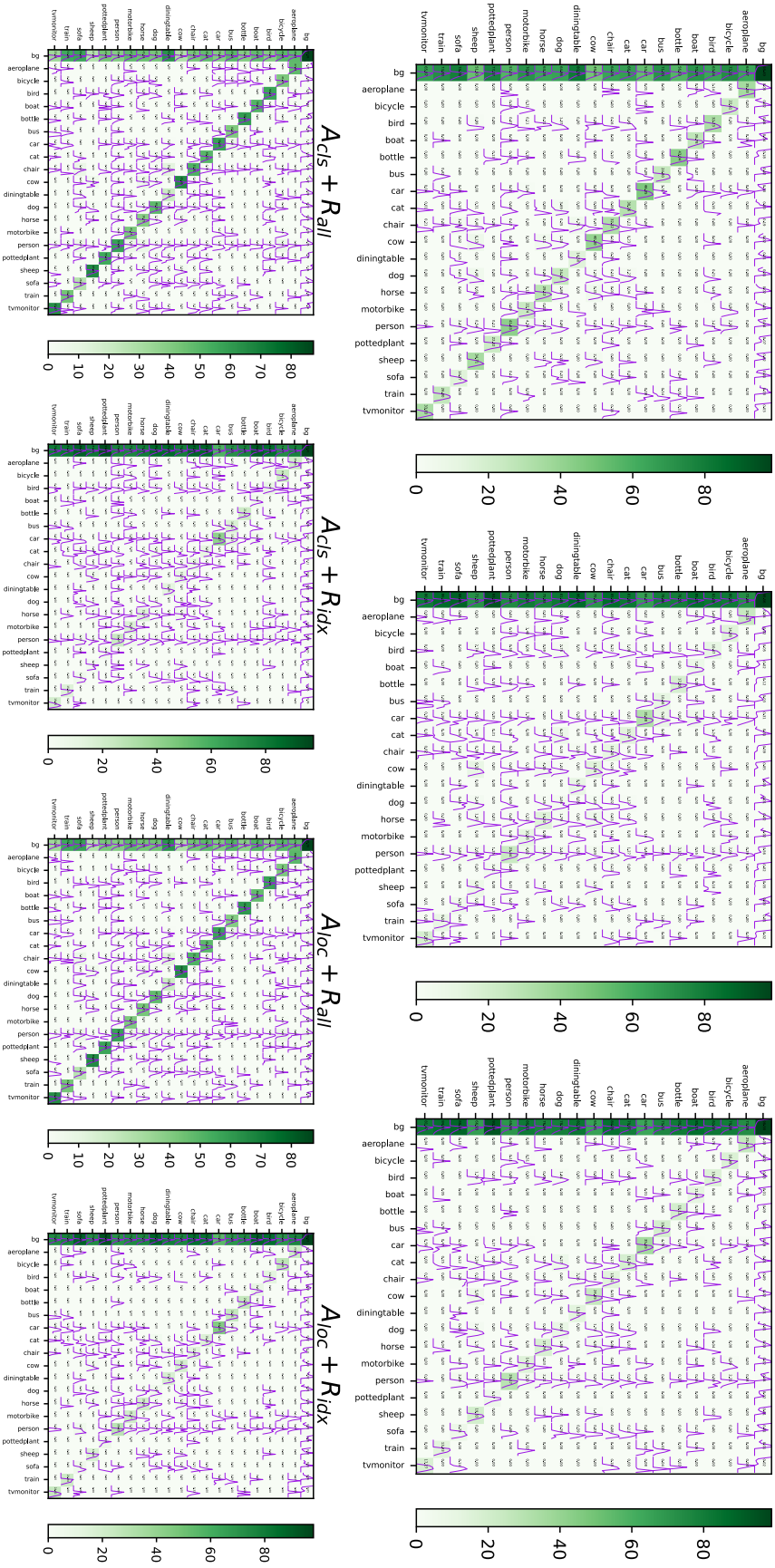

Figure X: DCM of the robust Faster RCNN Detector.

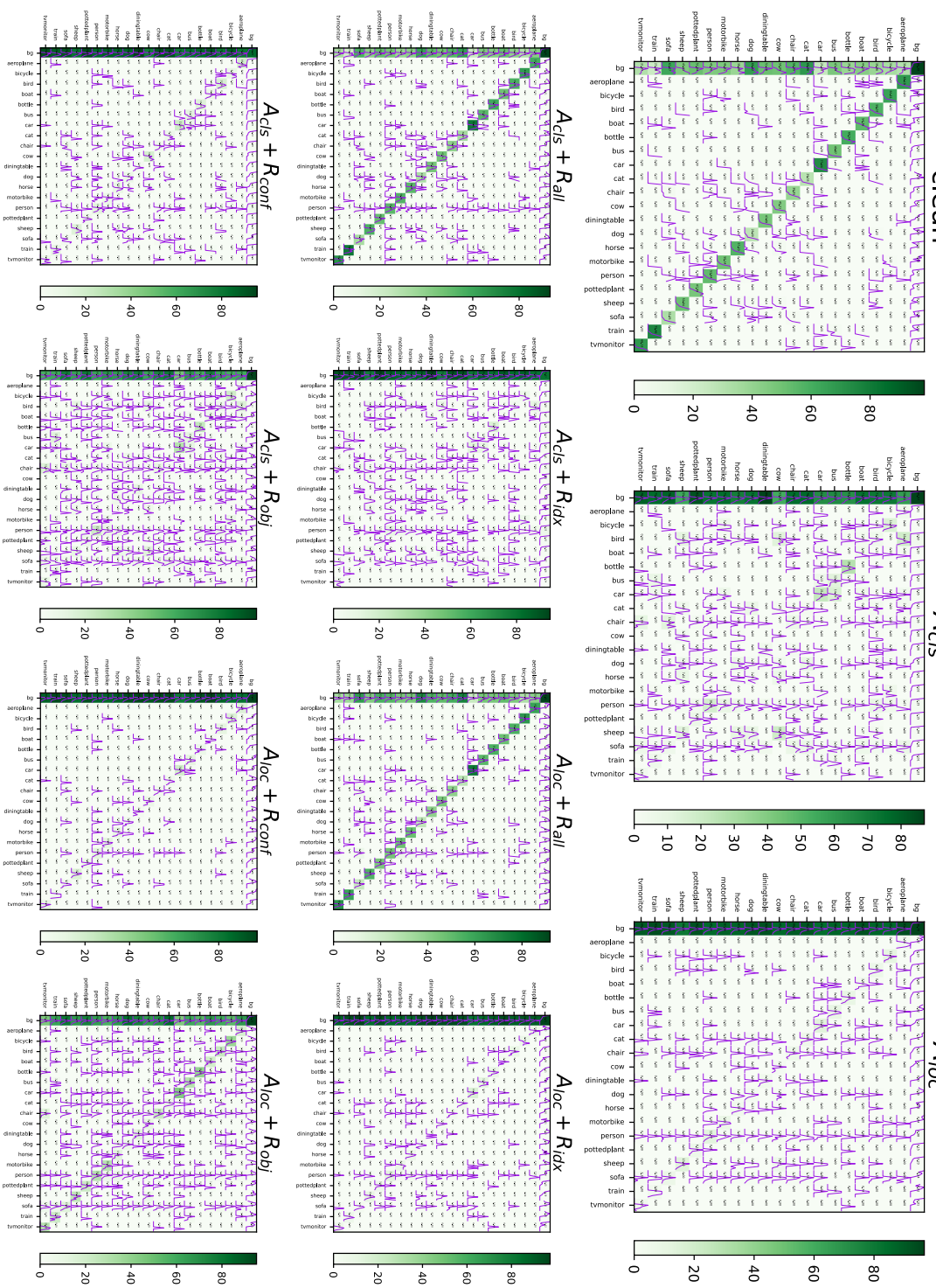

Figure XI: DCM of the standard YOLOX Detector.

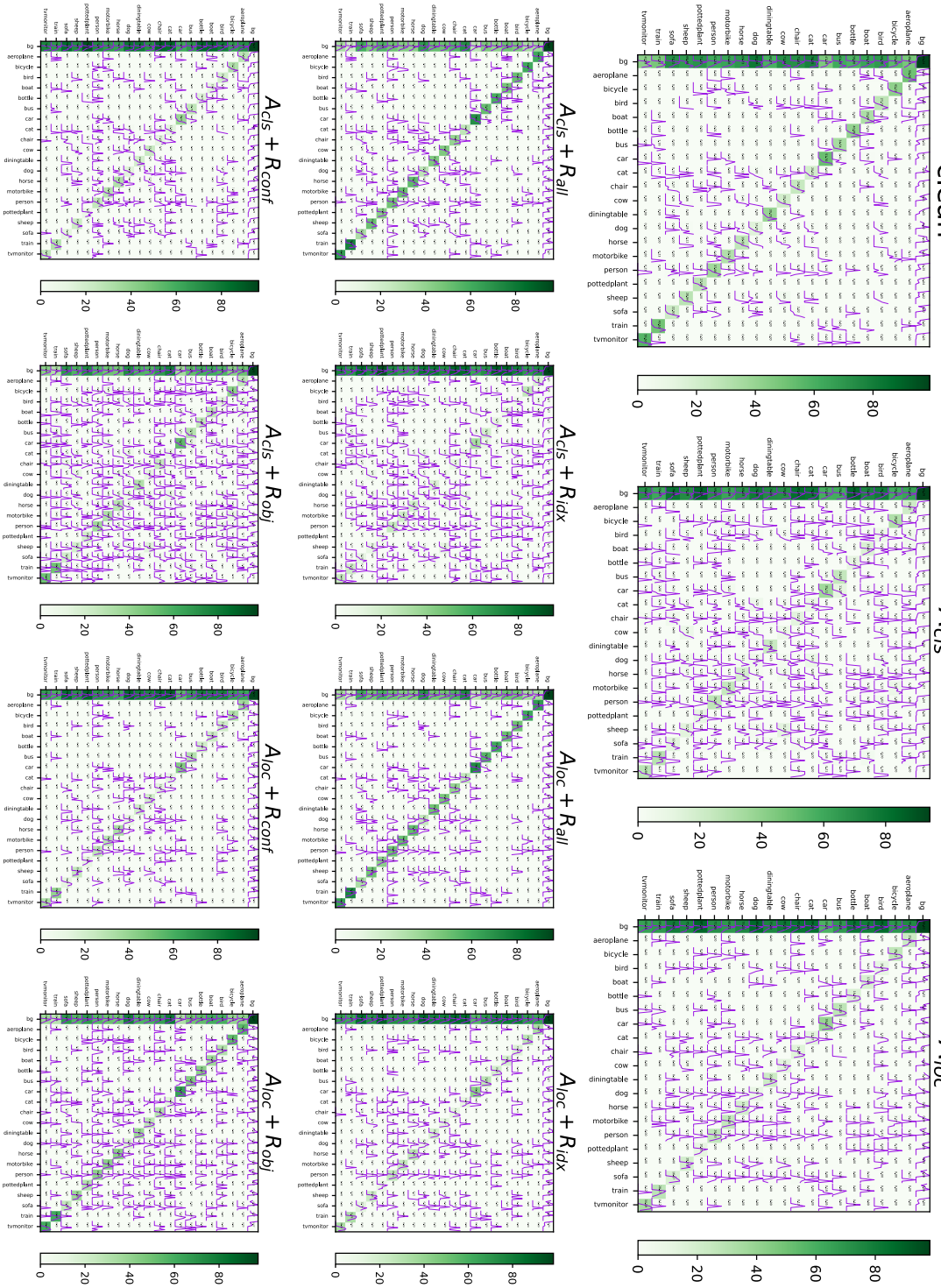

Figure XII: DCM of the robust YOLOX Detector.

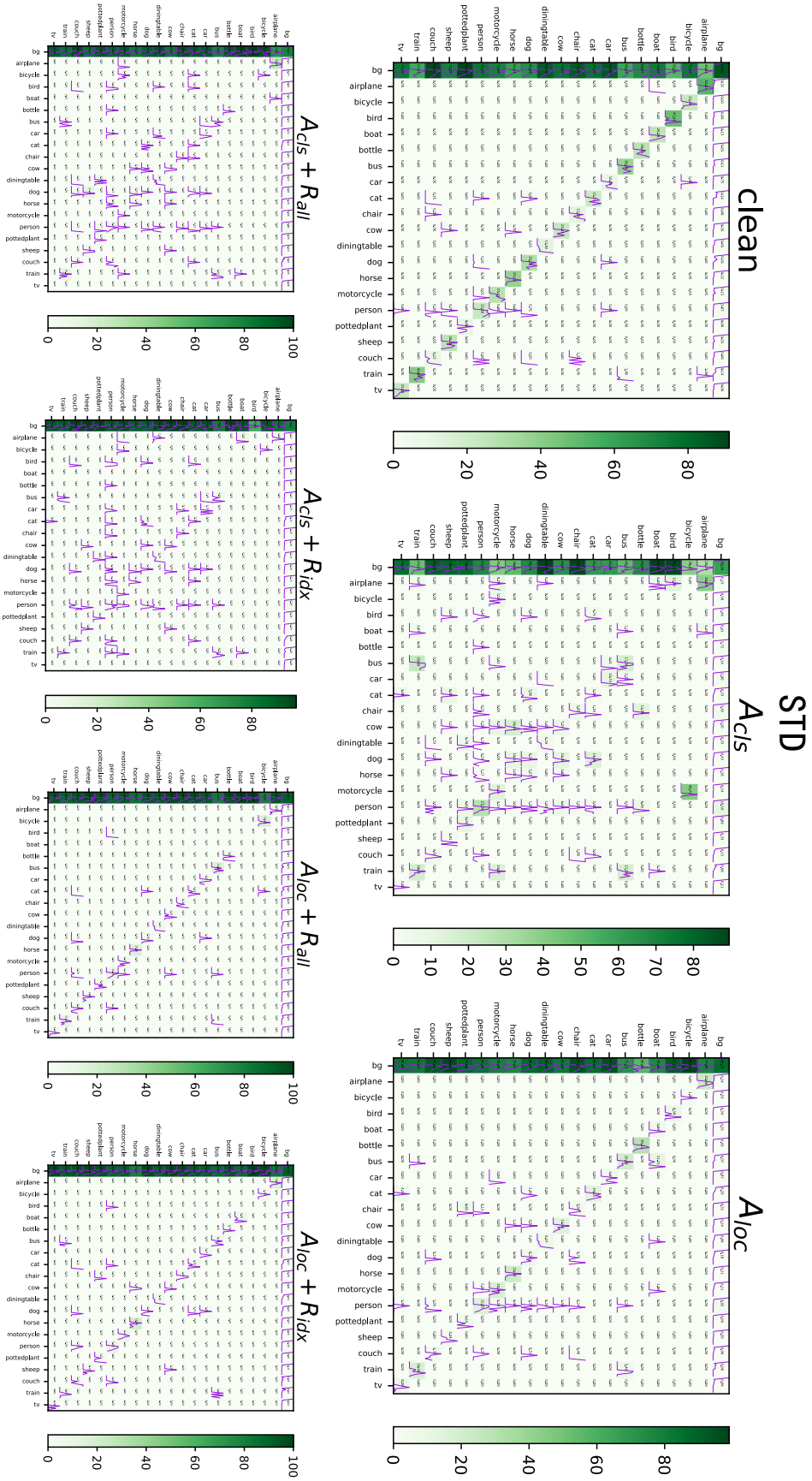

Figure XIII: DCM of the standard Deformable-DETR Detector.

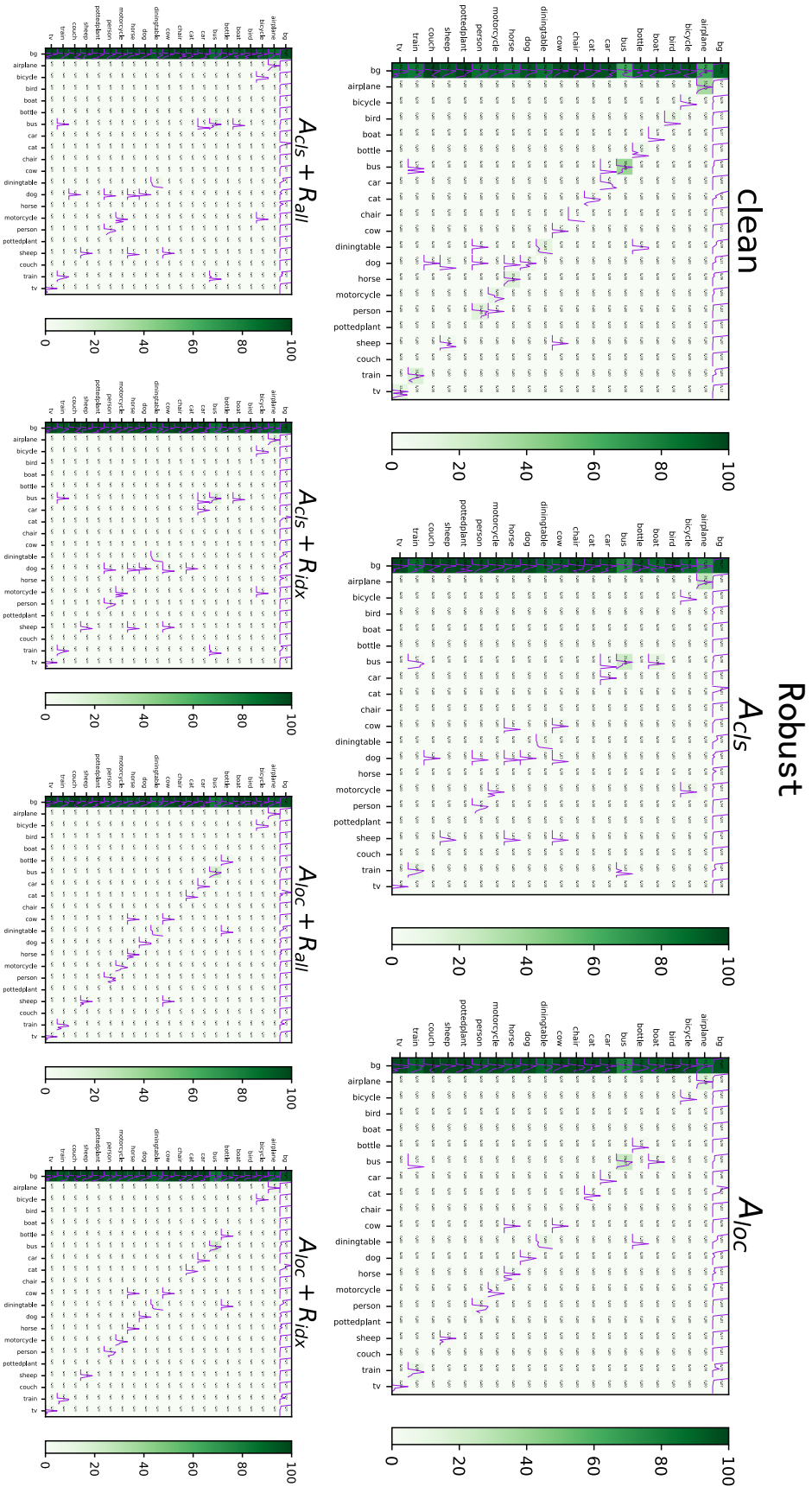

Figure XIV: DCM of the robust Deformable-DETR Detector.

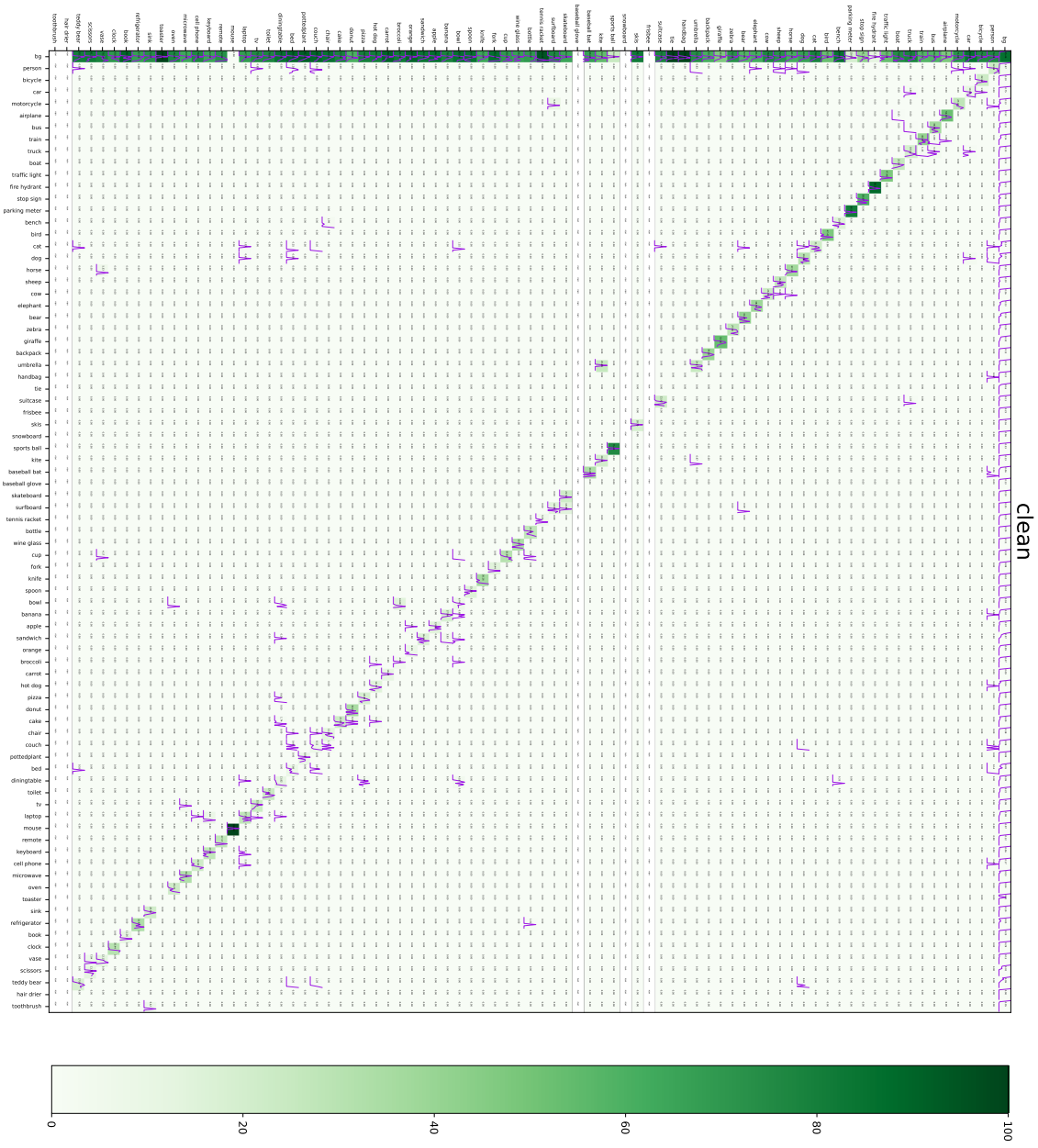

Figure XV: DCM of the standard Deformable-DETR Detector on all MS-COCO categories (clean).

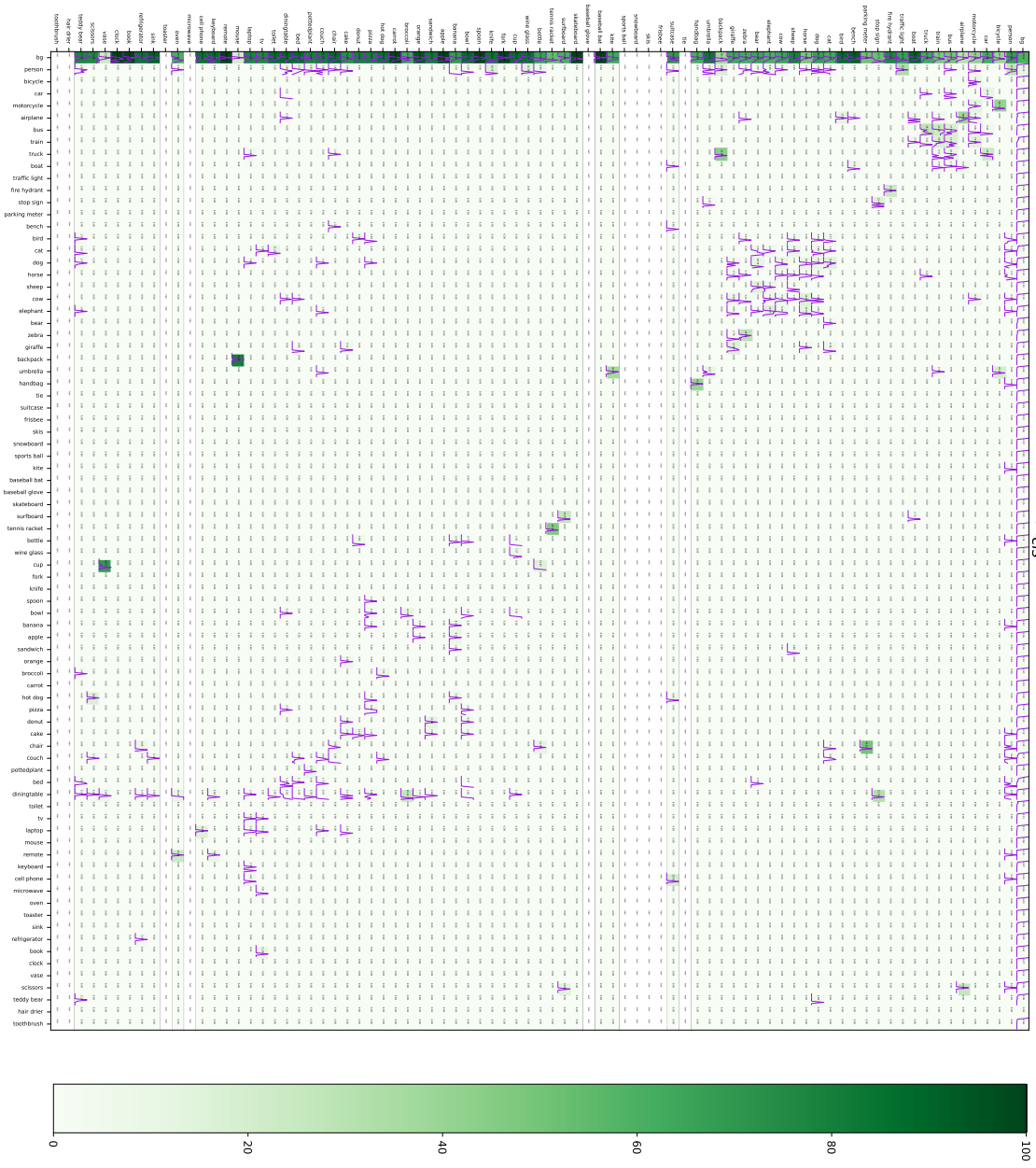

Figure XVI: DCM of the standard Deformable-DETR Detector on all MS-COCO categories (A_{cls}).

Figure XVII: DCM of the standard Deformable-DETR Detector on all MS-COCO categories ($A_{cls} + R_{all}$).

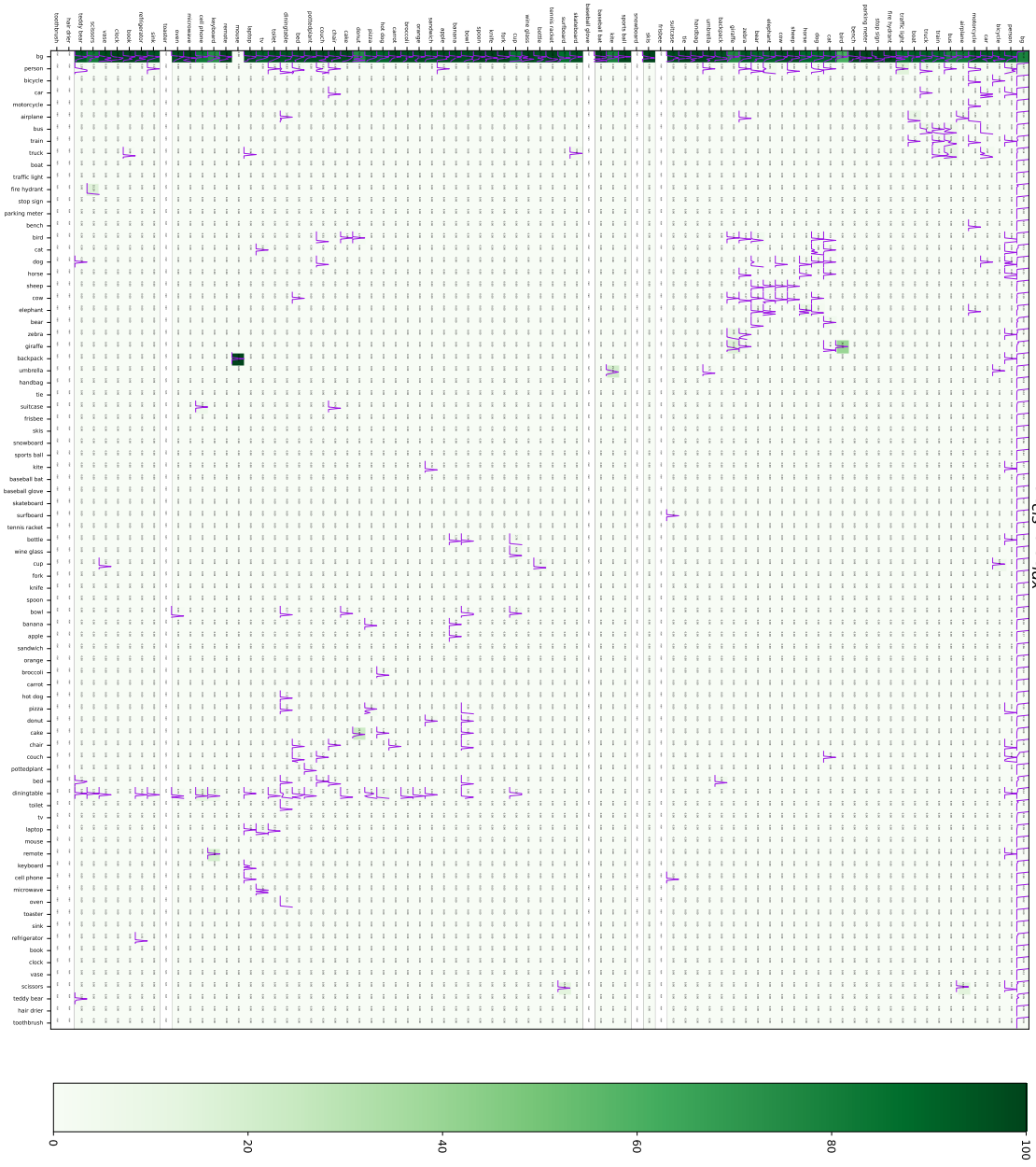

Figure XVIII: DCM of the standard Deformable-DETR Detector on all MS-COCO categories ($A_{cls} + R_{idx}$).

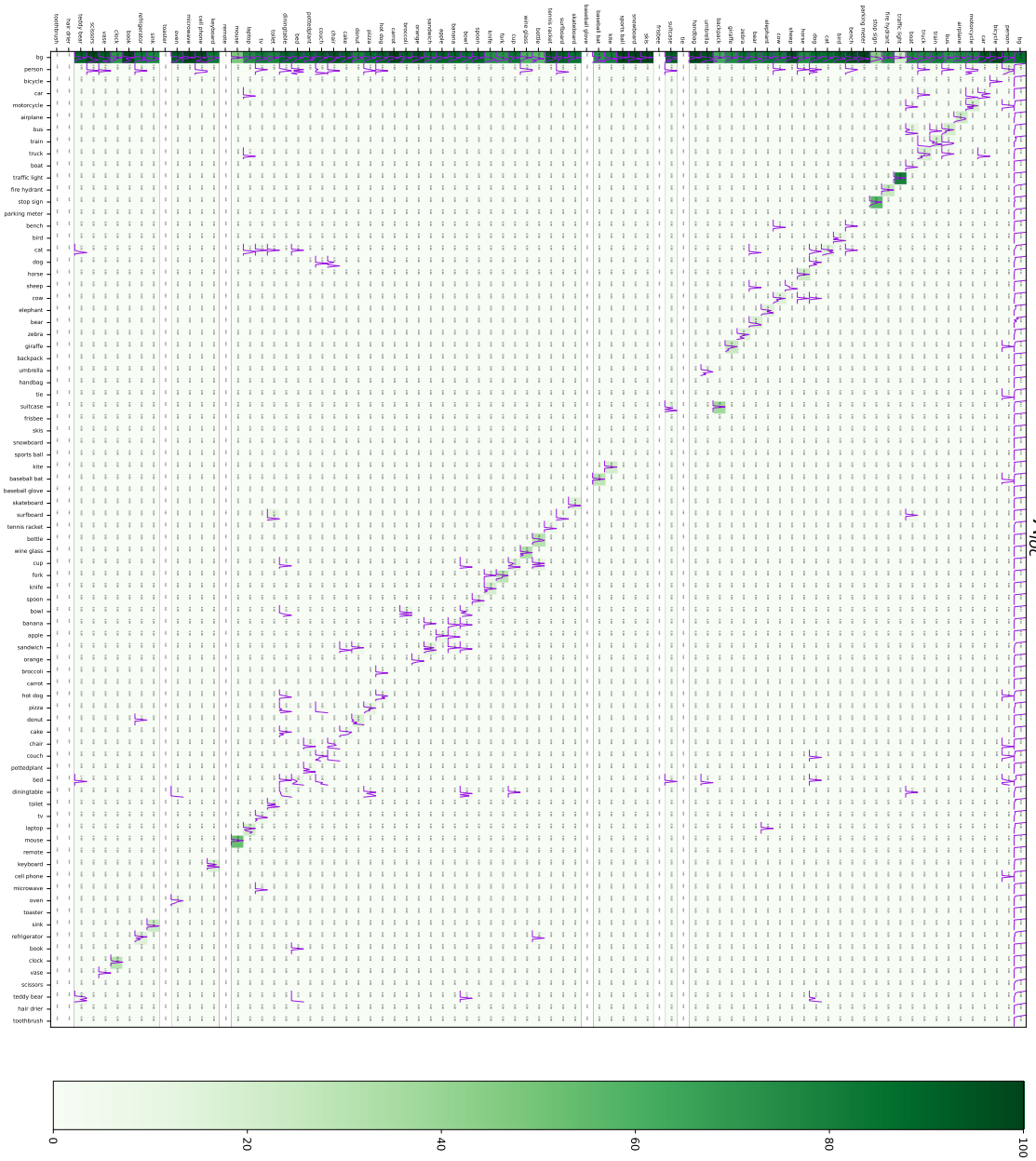

Figure XIX: DCM of the standard Deformable-DETR Detector on all MS-COCO categories (A_{loc}).

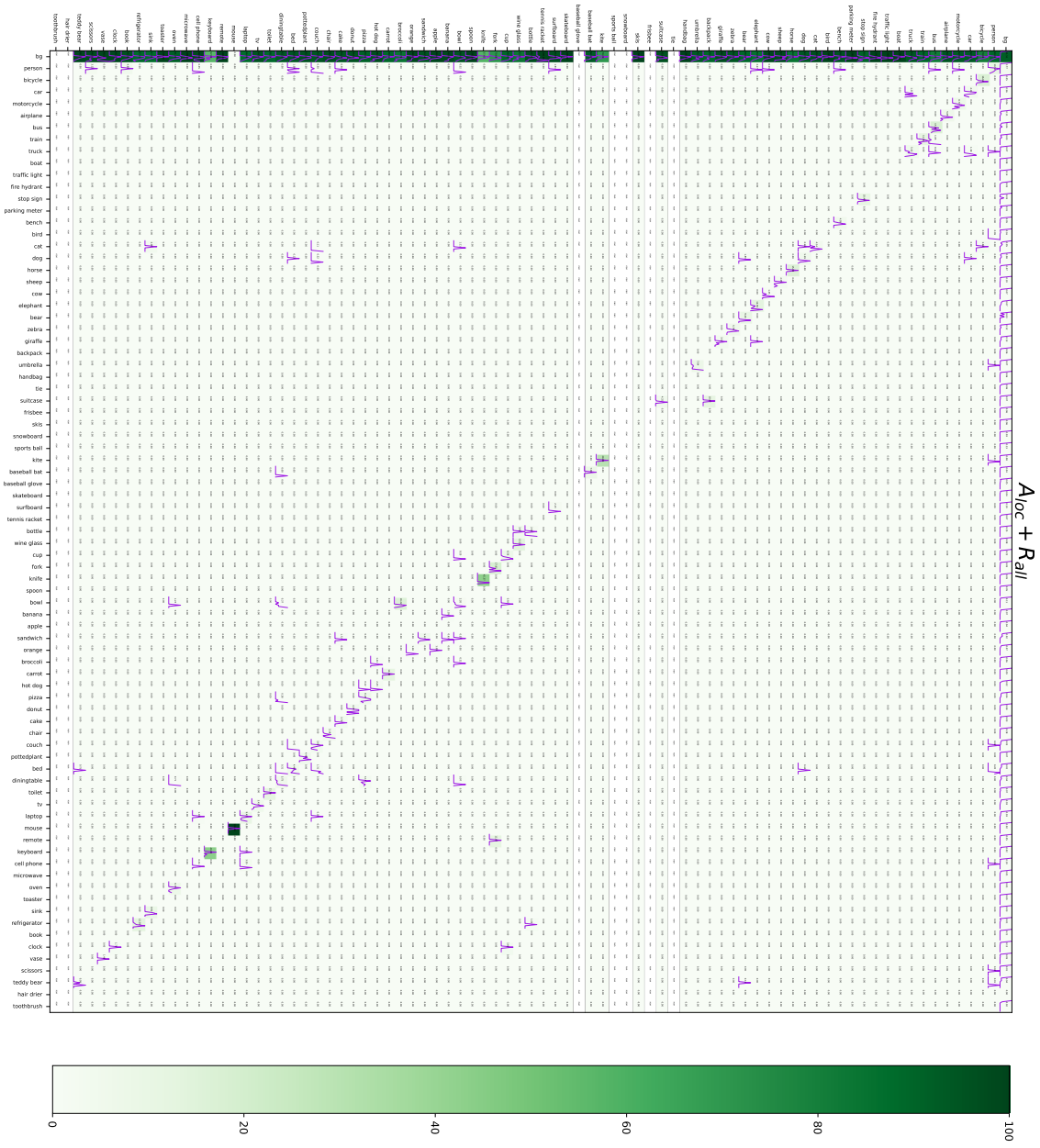

Figure XX: DCM of the standard Deformable-DETR Detector on all MS-COCO categories ($A_{loc} + R_{all}$).

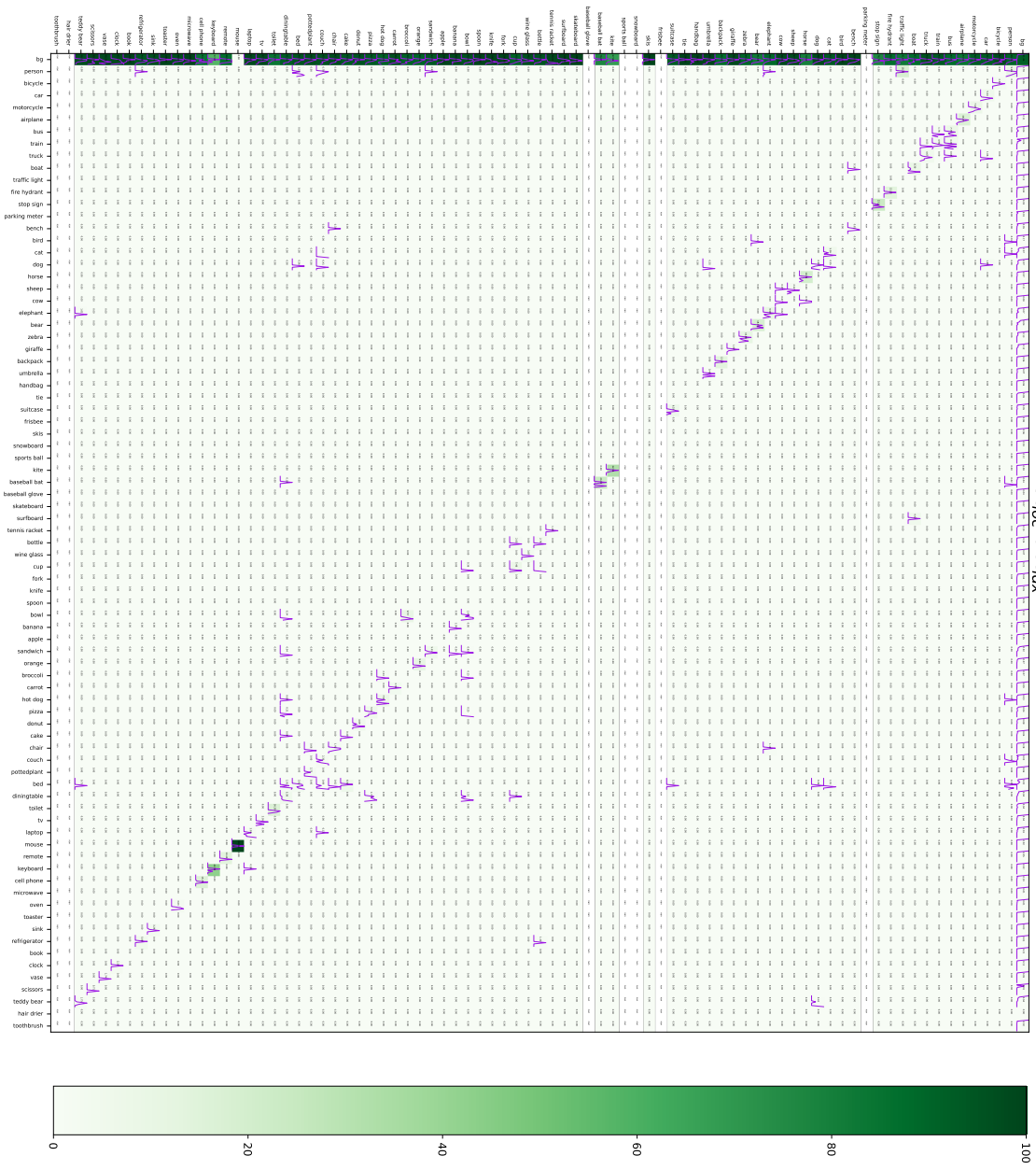

Figure XXI: DCM of the standard Deformable-DETR Detector on all MS-COCO categories ($A_{loc} + R_{idx}$).

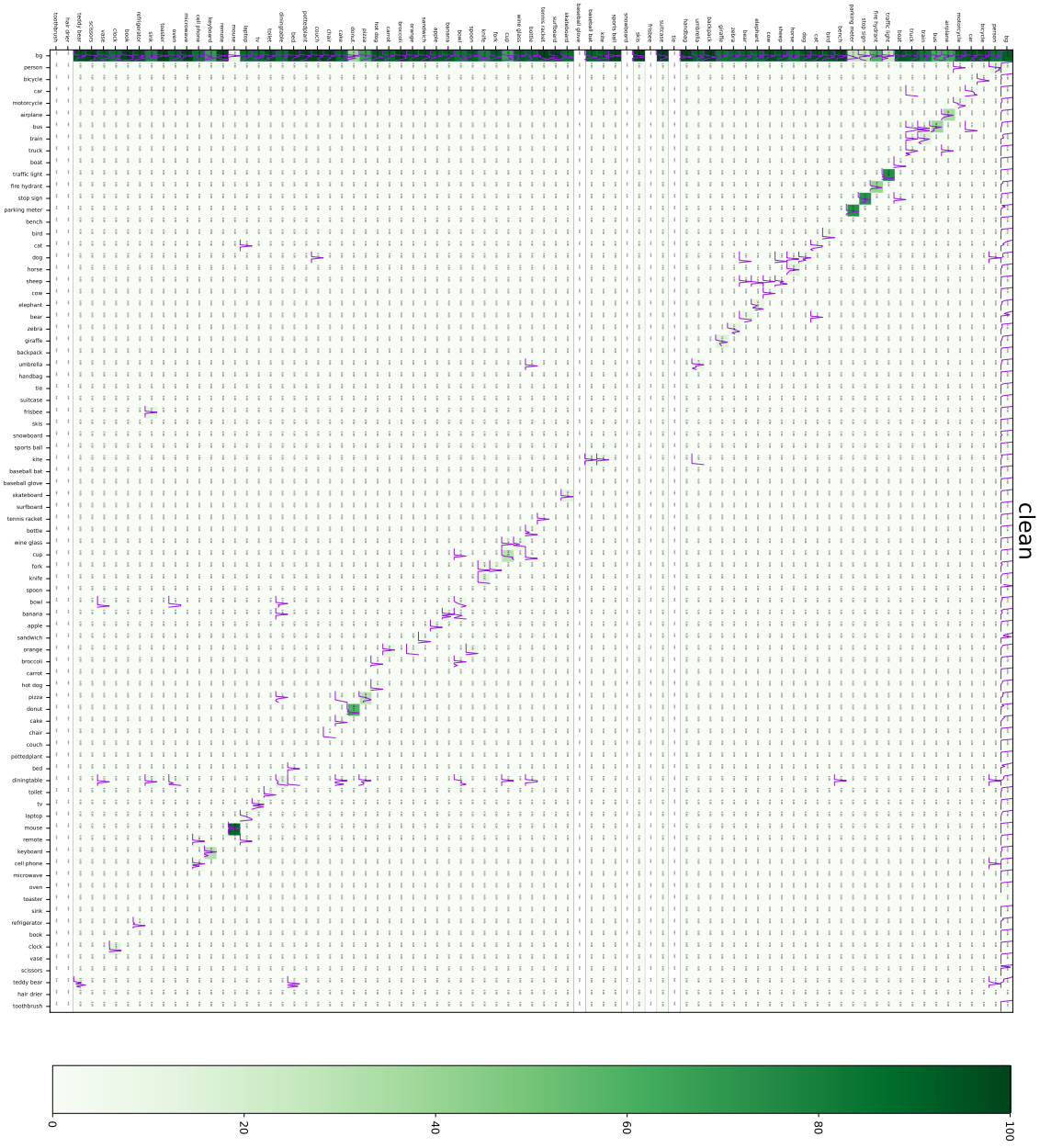

Figure XXII: DCM of the robust Deformable-DETR Detector on all MS-COCO categories (clean).

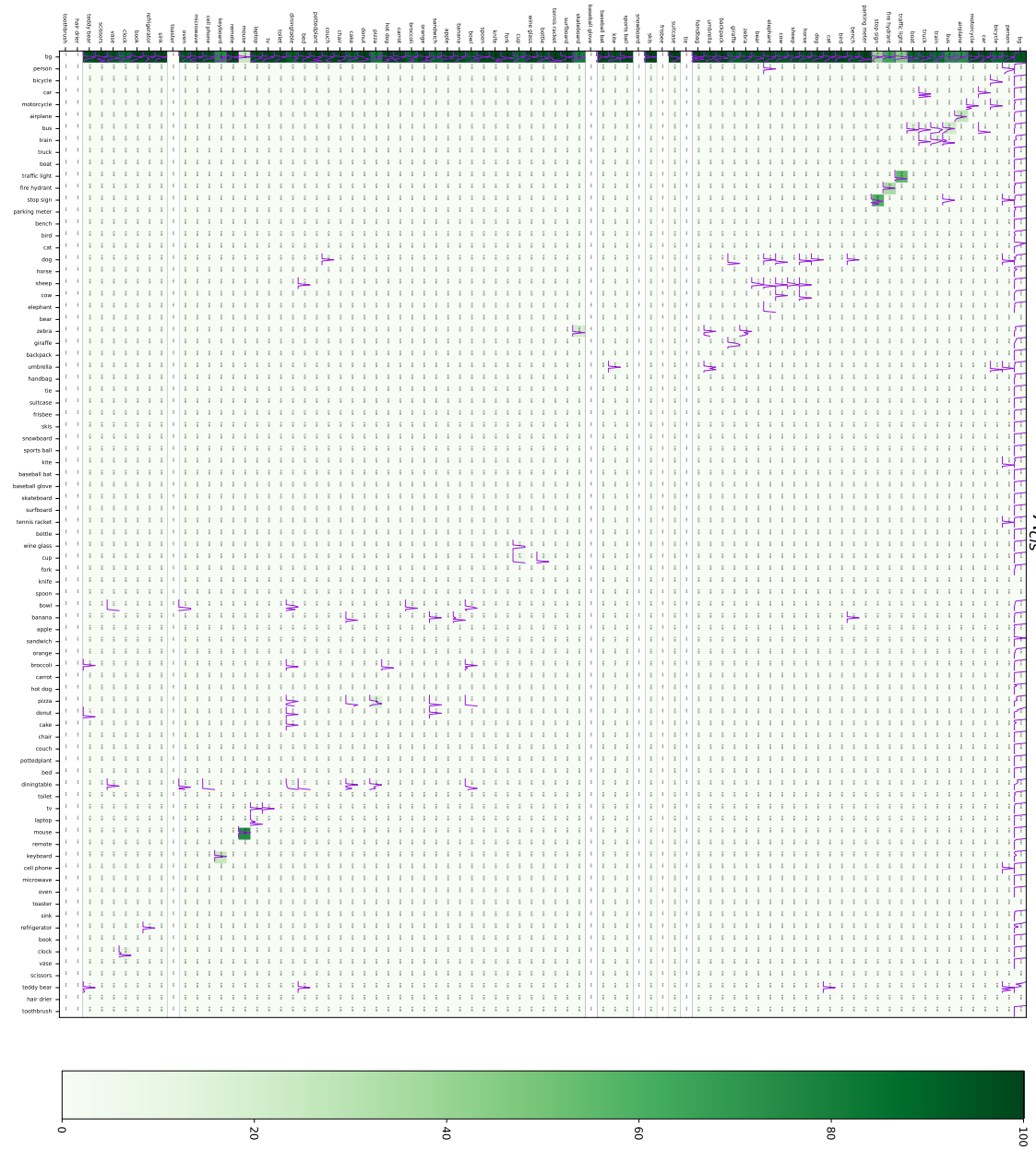

Figure XXIII: DCM of the robust Deformable-DETR Detector on all MS-COCO categories (A_{cls}).

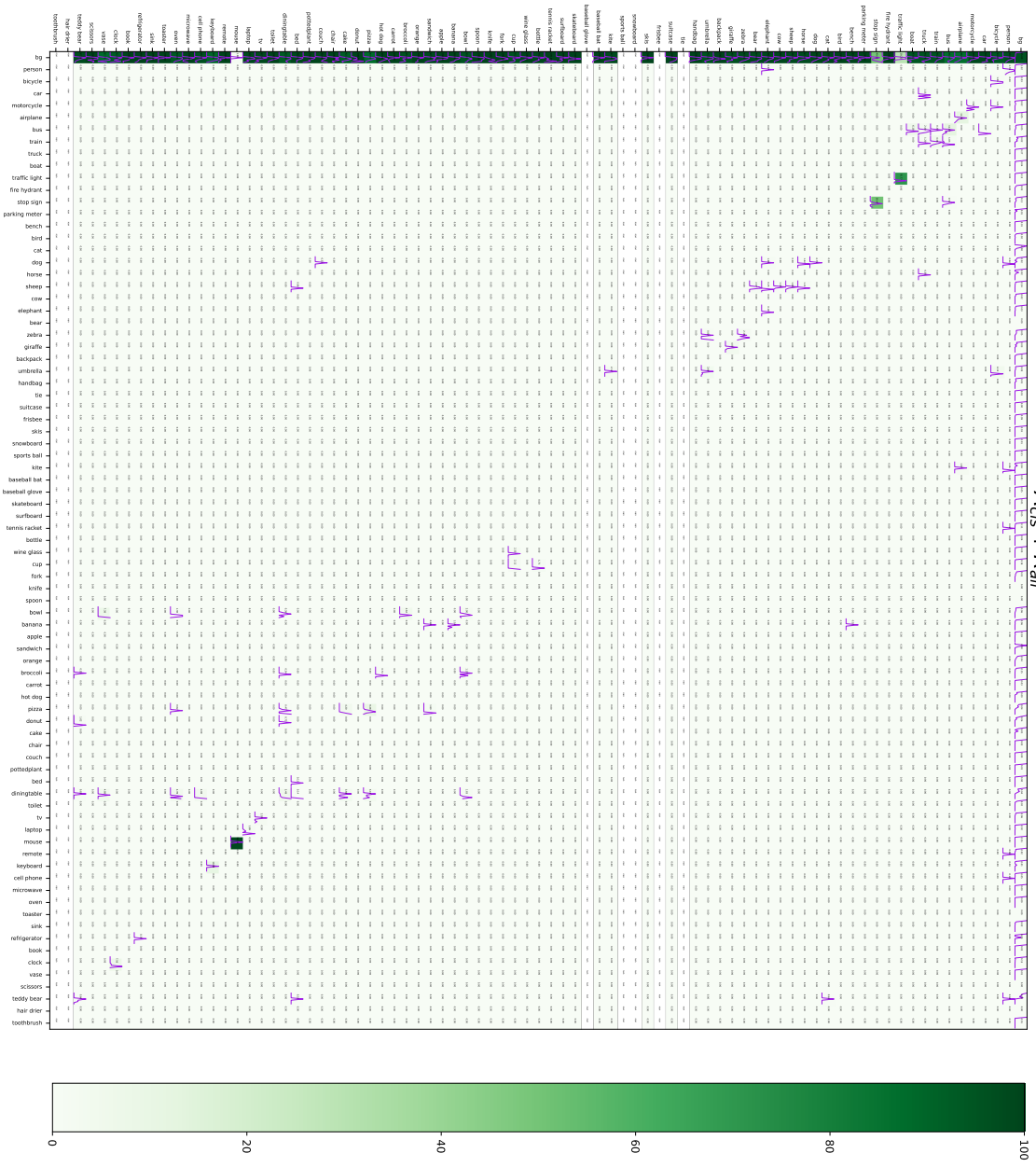

Figure XXIV: DCM of the robust Deformable-DETR Detector on all MS-COCO categories ($A_{cls} + R_{all}$).

Figure XXV: DCM of the robust Deformable-DETR Detector on all MS-COCO categories ($A_{cls} + R_{idx}$).

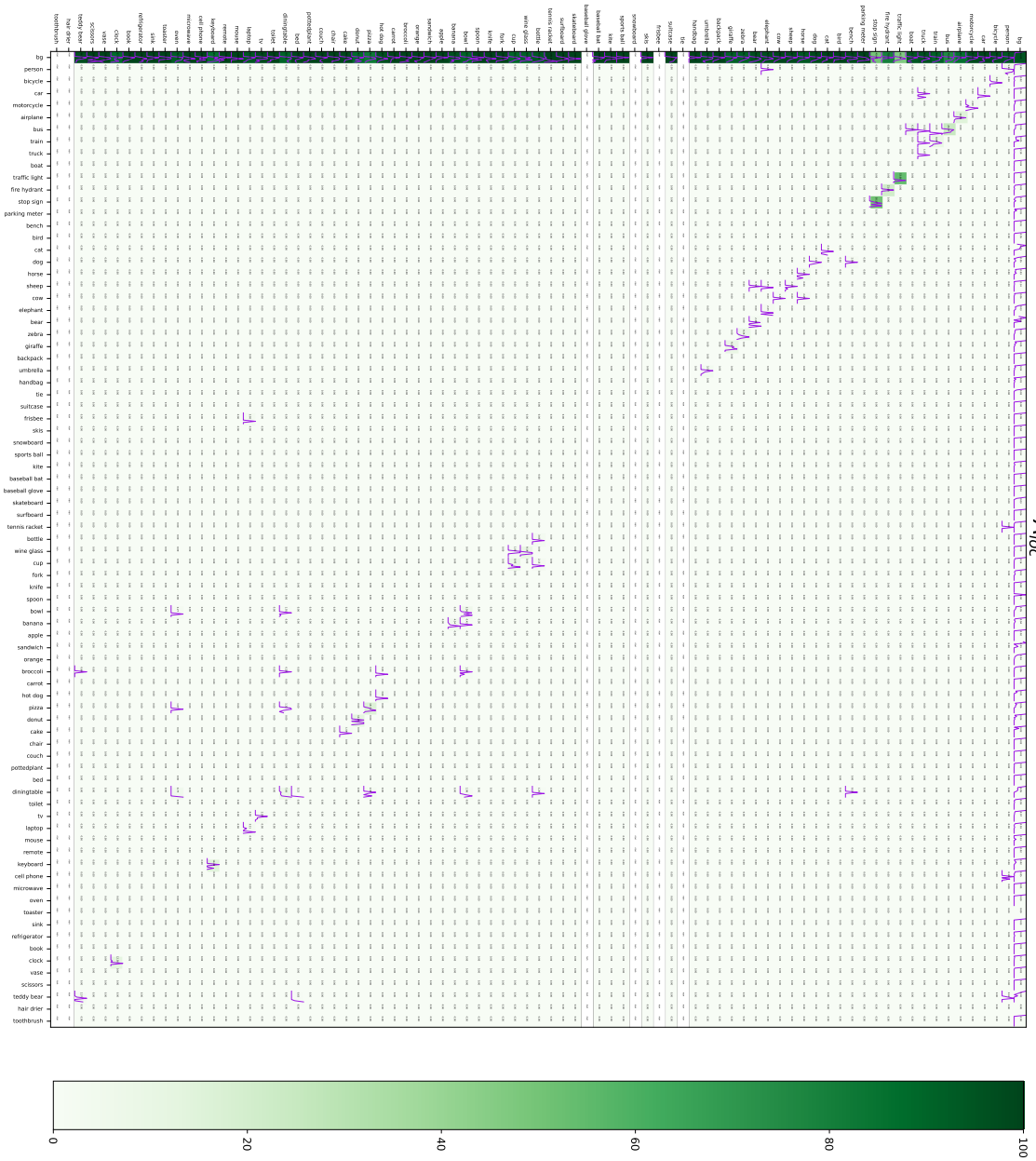

Figure XXVI: DCM of the robust Deformable-DETR Detector on all MS-COCO categories (A_{loc}).

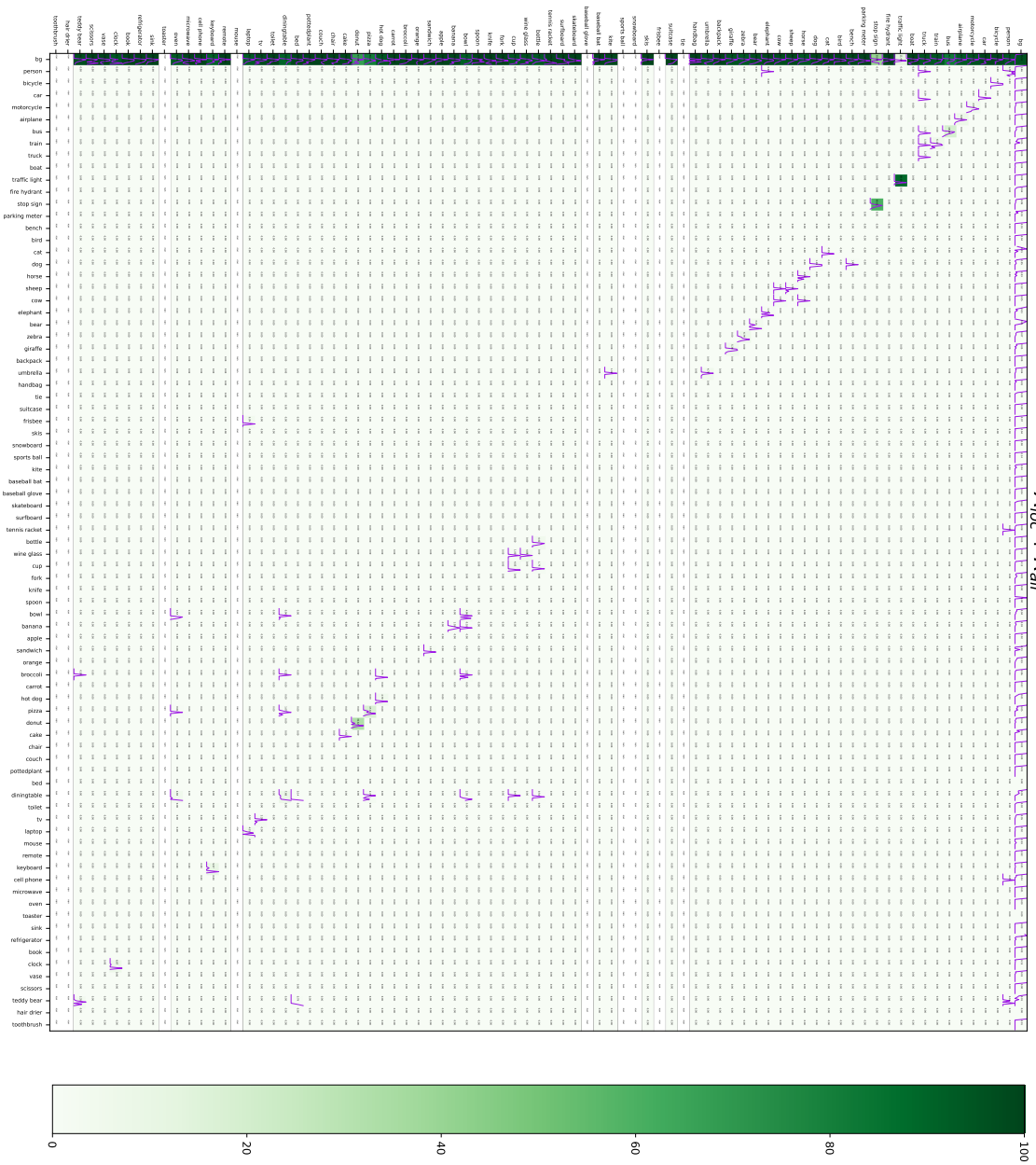

Figure XXVII: DCM of the robust Deformable-DETR Detector on all MS-COCO categories ($A_{loc} + R_{all}$).

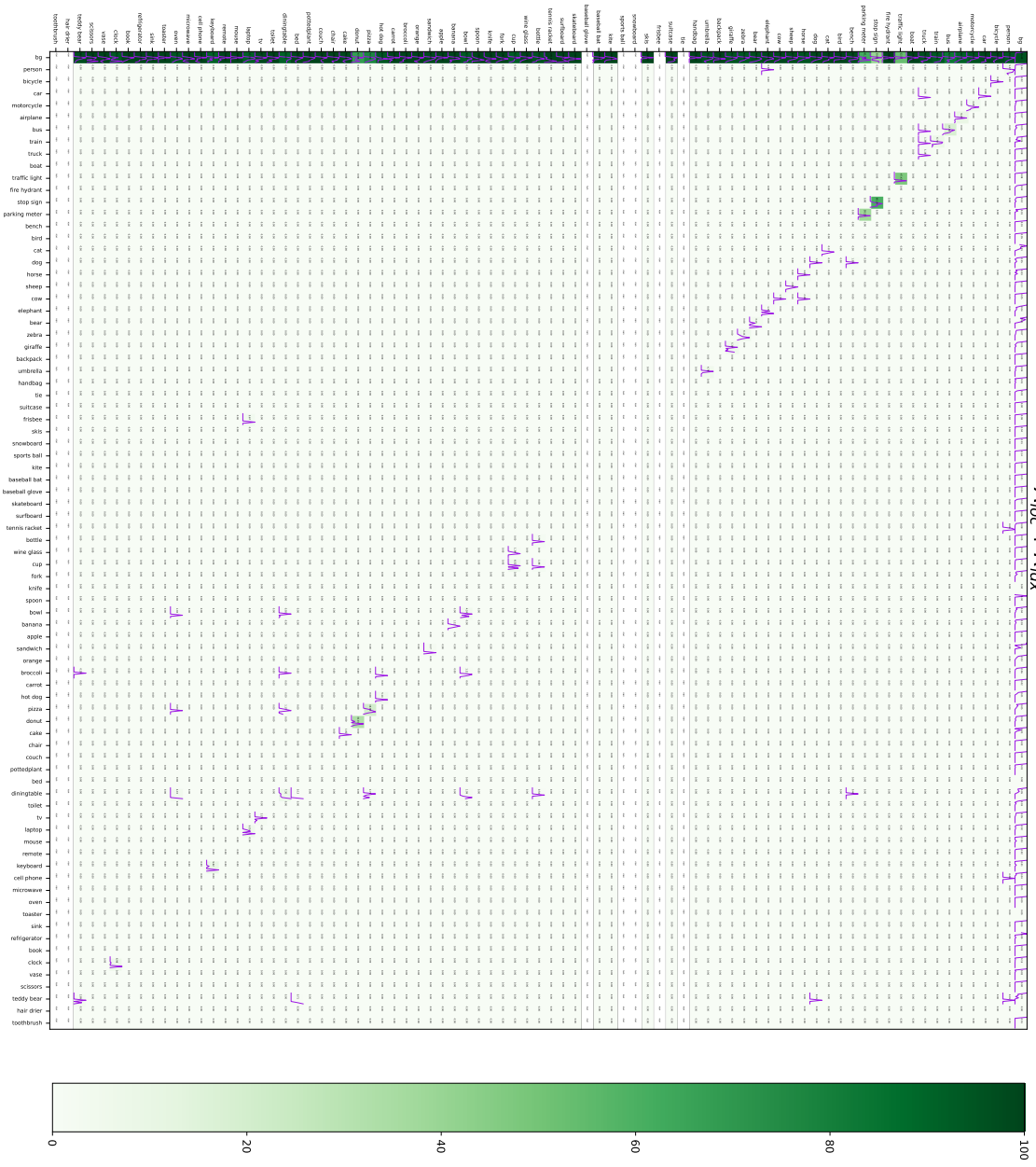

Figure XXVIII: DCM of the robust Deformable-DETR Detector on all MS-COCO categories ($A_{loc} + R_{idx}$).

REFERENCES

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.

Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

Pin-Chun Chen, Bo-Han Kung, and Jun-Cheng Chen. Class-aware robust adversarial training for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10420–10429, 2021.

Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng (Polo) Chau. Shapeshifter: Robust physical adversarial attack on faster R-CNN object detector. In *Machine Learning and Knowledge Discovery in Databases - European Conference (ECML)*, pp. 52–68, 2018.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, 2015.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1778–1787. Computer Vision Foundation / IEEE Computer Society, 2018.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, pp. 21–37, 2016.

Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Yiran Chen, and Hai Li. DPATCH: an adversarial patch attack on object detectors. In *Workshop on Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 13824–13833, 2019.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 91–99, 2015.

Sayantan Sarkar, Ankan Bansal, Upal Mahbub, and Rama Chellappa. UPSET and ANGRI : Breaking high performance image classifiers. *CoRR*, abs/1707.01159, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018.

Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 954–960, 2019.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. Adversarial examples for semantic segmentation and object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1378–1387, 2017.

Jiancheng Yang, Yangzhou Jiang, Xiaoyang Huang, Bingbing Ni, and Chenglong Zhao. Learning black-box attackers with transferable priors and query feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Haichao Zhang and Jianyu Wang. Towards adversarially robust object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 421–430, 2019.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan S. Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International Conference on Machine Learning (ICML)*, pp. 11278–11287, 2020.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021.

A Appendix

You may include other additional sections here.