南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

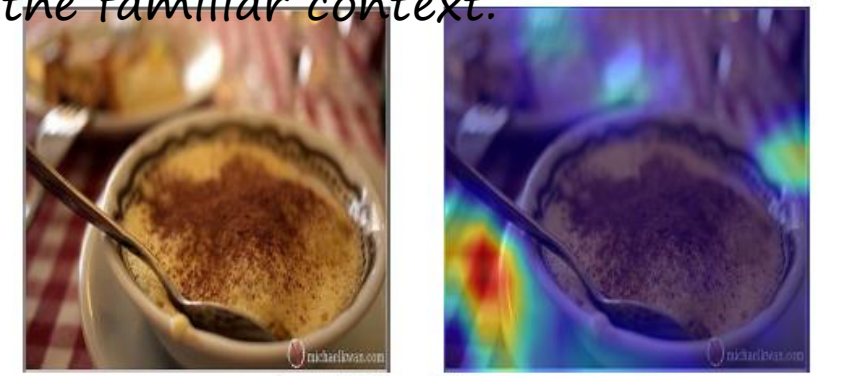
CAUSALITY COMPENSATED ATTENTION FOR CONTEXTUAL BIASED VISUAL RECOGNITION

Ruyang Liu [1] **Jingjia Huang** [2] **Ge Li** ✉[1] **Thomas H. Li**[1]
[1]School of Electronic and Computer Engineering, Peking University [2]ByteDance Inc
{ruyang@stu,geli@ece,thomas@}.pku.edu.cn huangjingjia@bytedance.com

ICLR 2023

Background

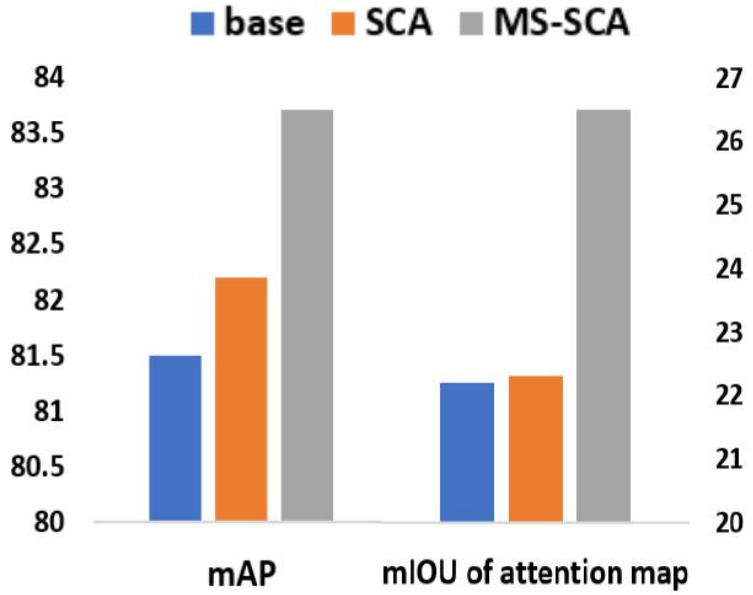
Helps the model make correct predictions even if the stressed region is wrong, but it fails in cases where the object is absent in the familiar context.



prob of "fork": 0.783

prob of "fork": 0.694

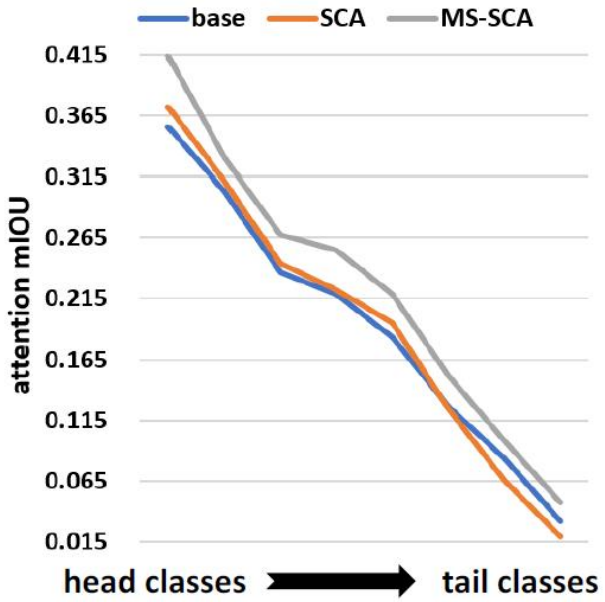
(a)Two qualitative results



(b)Comparison of mAP and mIOU

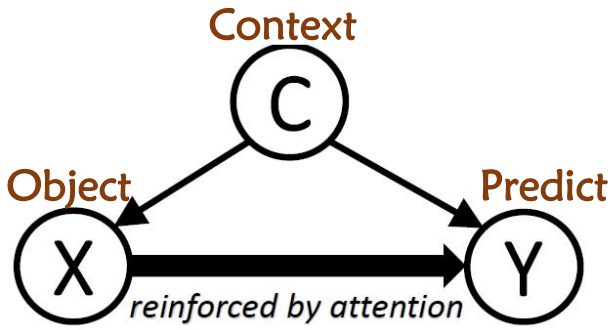
The attention does not capture the more accurate regions of targets than the baseline.

The attention mechanism more easily attends to the wrong context regions when the training samples are insufficient.

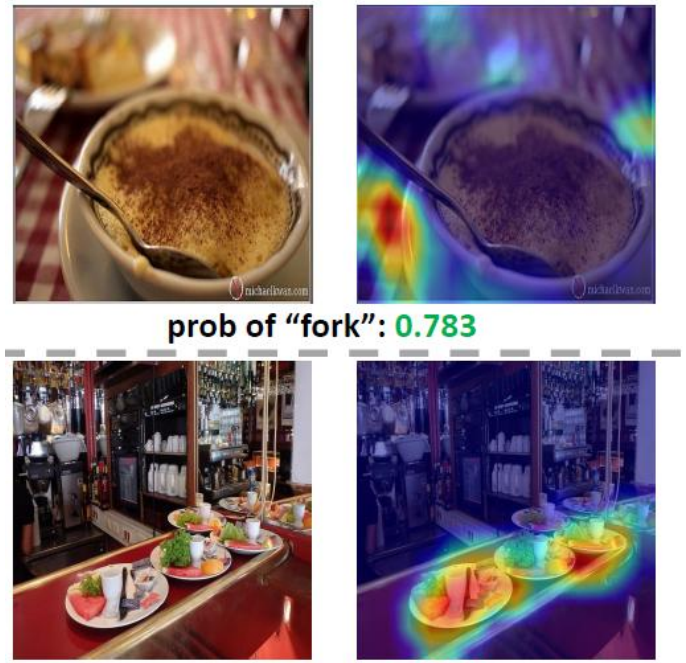


(c)mIOU from head to tail classes

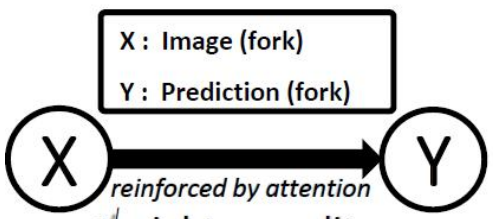
Background



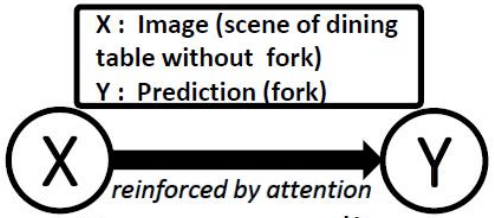
(a)Proposed causal graph



(a)Two qualitative results



√ right causality

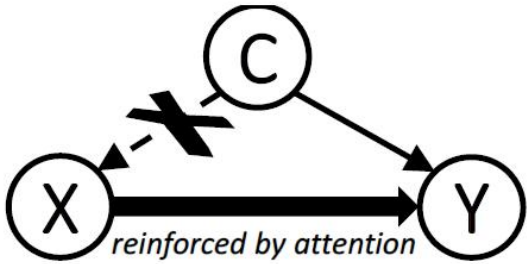


× wrong causality

Backdoor adjustment:

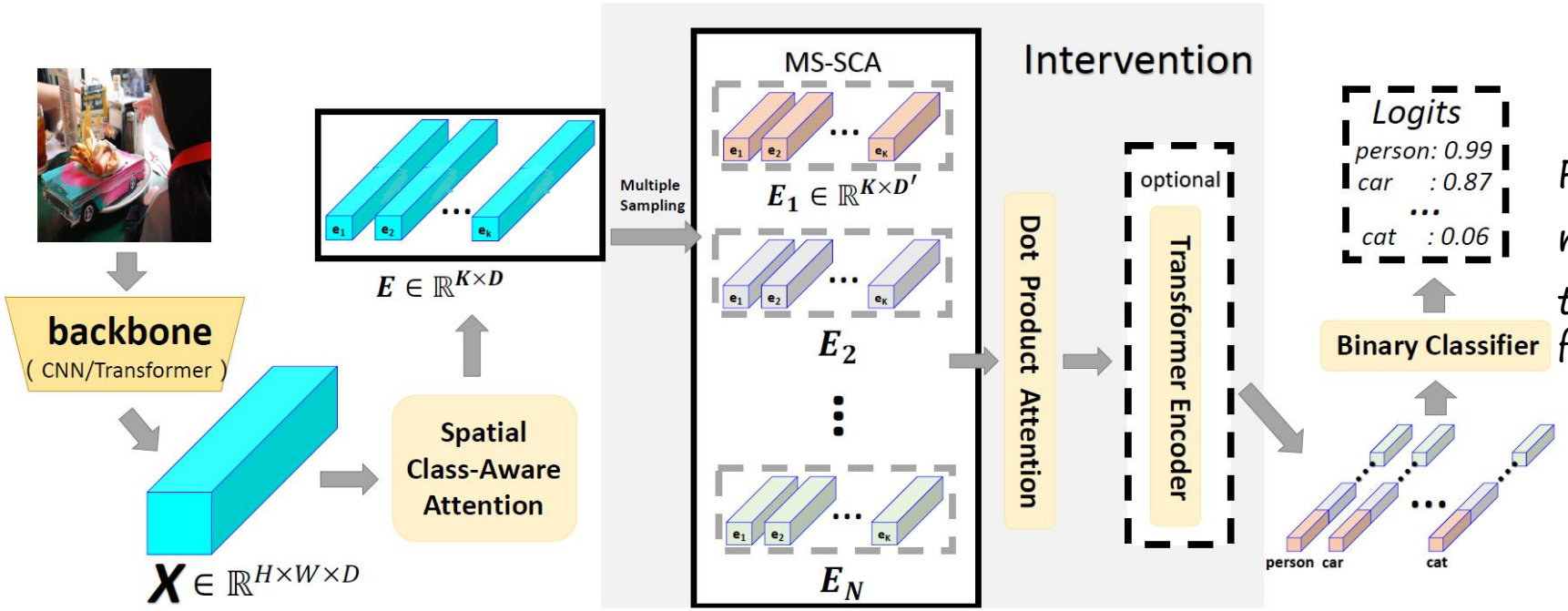
$do(X) = x'$, 表示将指向节点X的有向边全部切断，并将X的值固定为常数x'

$$\mathrm{P}(Y|\mathrm{do}(X)) = \sum_{c} \mathrm{P}(Y|X, C = c)\mathrm{P}(C = c)$$



(c) Causal intervention

Methods



Backdoor adjustment:

$$\mathrm{P}(Y|\mathrm{do}(X)) = \sum_c \mathrm{P}(Y|X, C=c)\mathrm{P}(C=c)$$

For each specific class k, its representation e_k is computed from the weighted average of the spatial feature in X:

$$\mathbf{e}_k = \sum_{i=1}^{H}\sum_{j=1}^{W} \mathrm{P}(Y=k|X=x_{i,j})x_{i,j}$$

$$\mathrm{P}(Y=k|\mathrm{do}(X=x)) = \sum_c \frac{\mathrm{P}(Y=k, X=x|C=c)\mathrm{P}(C=c)}{\mathrm{P}(X=x|C=c)} \quad \text{Propensity Score}$$

$$= \sum_c \frac{\mathrm{P}(Y=k, X=x, C=c)}{\mathrm{P}(X=x|C=c)},$$

$$\mathrm{P}(Y=k|\mathrm{do}(X=x)) \approx \sum_{n=1}^{N} \frac{\mathrm{P}(Y=k, X=x^n, C=c)}{\mathrm{P}(X=x^n|C=c)}$$

$$\mathrm{P}(Y=k|\mathrm{do}(X=x)) = \sum_{n=1}^{N} \frac{\mathrm{P}(Y=k|X=x^n)\mathrm{P}(X=x^n)}{\mathrm{P}(X=x^n|C=c)} = \sum_{n=1}^{N} \frac{\mathrm{Sigmoid}(w_k e_k^n)\mathrm{P}(e_k^n)}{\mathrm{P}(X=e_k^n|C=c)}$$

The Propensity Score in the classification model following Rubin's theory:

$$\mathrm{P}(X=e_k^n|C=c) \rightarrow (\|w_k\|_2 \cdot \|e_k^n\|_2) + (\gamma \cdot \|e_k^n\|_2)$$

The ultimate intervention:

$$\mathrm{P}(Y=k|\mathrm{do}(X=x)) = \sum_{n=1}^{N} \frac{\mathrm{Sigmoid}(w_k e_k^n)}{\|w_k\|_2 + \gamma}\mathrm{P}(e_k^n)$$

Methods

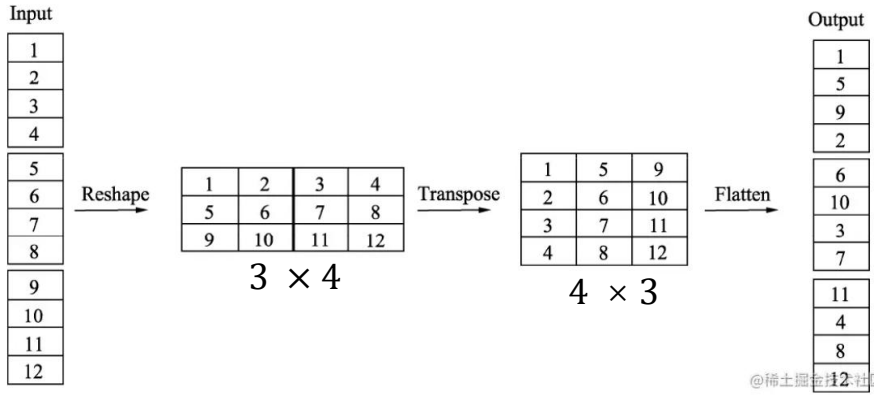
$$\mathbf{e}_k = \sum_{i=1}^{H}\sum_{j=1}^{W} \mathrm{P}(Y=k|X=x_{i,j})x_{i,j}$$

$$\mathrm{P}(Y=k|\mathrm{do}(X=x)) = \sum_{n=1}^{N} \frac{\mathrm{Sigmoid}(w_k e_k^n)}{\|w_k\|_2 + \gamma} \mathrm{P}(e_k^n)$$

- Random sampling
- Multi-head

  divide the channel into N groups
  and take each group as a sample

- Channel-shuffle



- Scaled Dot-Product Attention (DPA)

$$\mathbf{E} \in \mathbb{R}^{B\times N\times K\times D'} \Rightarrow \mathbf{E}^s \in \mathbb{R}^{N\times(B*K)\times D'}$$

$$\mathbf{E}^{s'} = \mathrm{softmax}(\frac{(W_q\mathbf{E}^s)(W_k\mathbf{E}^s)^T}{\sqrt{D'}})(W_v\mathbf{E}^s) \quad \text{Light}$$

$$\mathbf{E}^{s+1} = \max(0, \mathbf{E}^{s'}W_1 + b_1)W_2 + b_2 \quad \text{Heavy}$$

- Average

  assign $\mathrm{P}(e_k^n)$ as $1/N$

- learnable re-weight parameters

Experiments

Methods	Resolutions	mAP	All		Top3	
			CF1	OF1	CF1	OF1
ResNet-101 (He et al., 2016)	448 * 448	81.5	76.3	80.0	73.5	76.0
ML-GCN (Chen et al., 2019c)	448 * 448	83.0	78.0	80.3	74.2	76.3
MS-CMA (You et al., 2020)	448 * 448	83.8	78.4	81.0	**74.9**	77.1
CSRA (Zhu & Wu, 2021)	448 * 448	83.5	77.9	80.3	74.4	76.5
IDA-R101(L)	448 * 448	84.3	78.5	**81.1**	73.6	77.3
IDA-R101(H)	448 * 448	**84.8**	**78.7**	80.9	73.9	**77.4**
SSGRL (Chen et al., 2019b)	576 * 576	83.8	76.8	79.7	72.7	76.2
C-Trans (Lanchantin et al., 2021)	576 * 576	85.1	79.9	81.7	76.0	77.6
ADD-GCN (Ye et al., 2020)	576 * 576	85.2	80.1	82.0	75.8	77.9
CCD (Liu et al., 2022)	576 * 576	85.3	80.2	82.1	76.0	77.9
TDRG (Zhao et al., 2021)	576 * 576	86.0	**80.4**	82.4	76.2	78.1
IDA-R101(L)	576 * 576	85.5	79.8	82.3	75.2	78.1
IDA-R101(H)	576 * 576	**86.3**	**80.4**	**82.5**	**76.4**	**78.2**
Swin-Base (Liu et al., 2021)	384 * 384	88.4	82.2	84.0	77.2	79.9
Swin-Large (Liu et al., 2021)	384 * 384	89.3	83.6	85.6	78.1	80.8
IDA-SwinB(H)	384 * 384	89.3	83.7	85.1	78.0	80.4
IDA-SwinL(H)	384 * 384	**90.3**	**84.7**	**85.9**	**79.0**	**81.1**

The first block shows that our light model has an obvious advantage over other methods with comparable computation.
For heavyweight models, other models universally adopt 3 or more layers of transformer or GNN, while ours only has two layers but outperforms all of them, confirming the trade-off between computation and performance in our method is more efficient.

Methods	Detector	MS-COCO mAP@.5	MS-COCO mAP@.75	MS-COCO bbox-mAP	VOC07 mAP@.5
Baseline	Faster-RCNN-ROIAlign-R50	58.1	40.5	37.4	80.5
IDA (L)	Faster-RCNN-ROIAlign-R50	**59.3**	**42.0**	**38.7**	**83.8**
Baseline	RetinaNet-R50	52.5	36.6	34.3	77.8
IDA (L)	RetinaNet-R50	**55.8**	**38.6**	**36.4**	**80.5**
Baseline	DeTR-DC5-R101	64.7	47.7	44.9	-
IDA (L)	DeTR-DC5-R101	**65.7**	**49.2**	**46.1**	-

Experiments

Components				mAP	
SCA	Multi-Sampling	DPA	Trans	448*448	576*576
				81.4	82.7
✓				81.9	83.0
	✓			82.1	83.0
✓	✓			83.6	84.8
	✓	✓		82.2	83.1
✓	✓	✓		84.3	85.5
✓	✓	✓	✓	**84.8**	**86.3**

$$\mathrm{P}(Y=k|\mathrm{do}(X=x))=\sum_{n=1}^{N}\frac{\mathrm{Sigmoid}(w_k e_k^n)}{\|w_k\|_2+\gamma}\mathrm{P}(e_k^n)$$

Sampling	$\mathrm{P}(e_k^n)$	mAP
baseline SCA	-	81.9
Random Sampling	Average	82.9
Multi-Head (2)	Average	82.8
Multi-Head (4)	Average	83.2
Multi-Head (8)	Average	83.4
Channel-shuffle	Average	**83.6**
Channel-shuffle	Average	83.6
Channel-shuffle	Parameter	83.9
Channel-shuffle	DPA	**84.3**

we sample and build a contextual biased subset from COCO. We sample half of the training set on MSCOCO in this way to form an OOD test set, and train on the rest training set.

1. select six classes of worst performance (e.g., toothbrush) on the original test set.
2. rank the top5 co-exist classes with these target classes (e.g., toilet for toothbrush).
3. choose target classes arising in no frequently co-exist classes as positive samples, and the familiar context without target classes as negative samples.

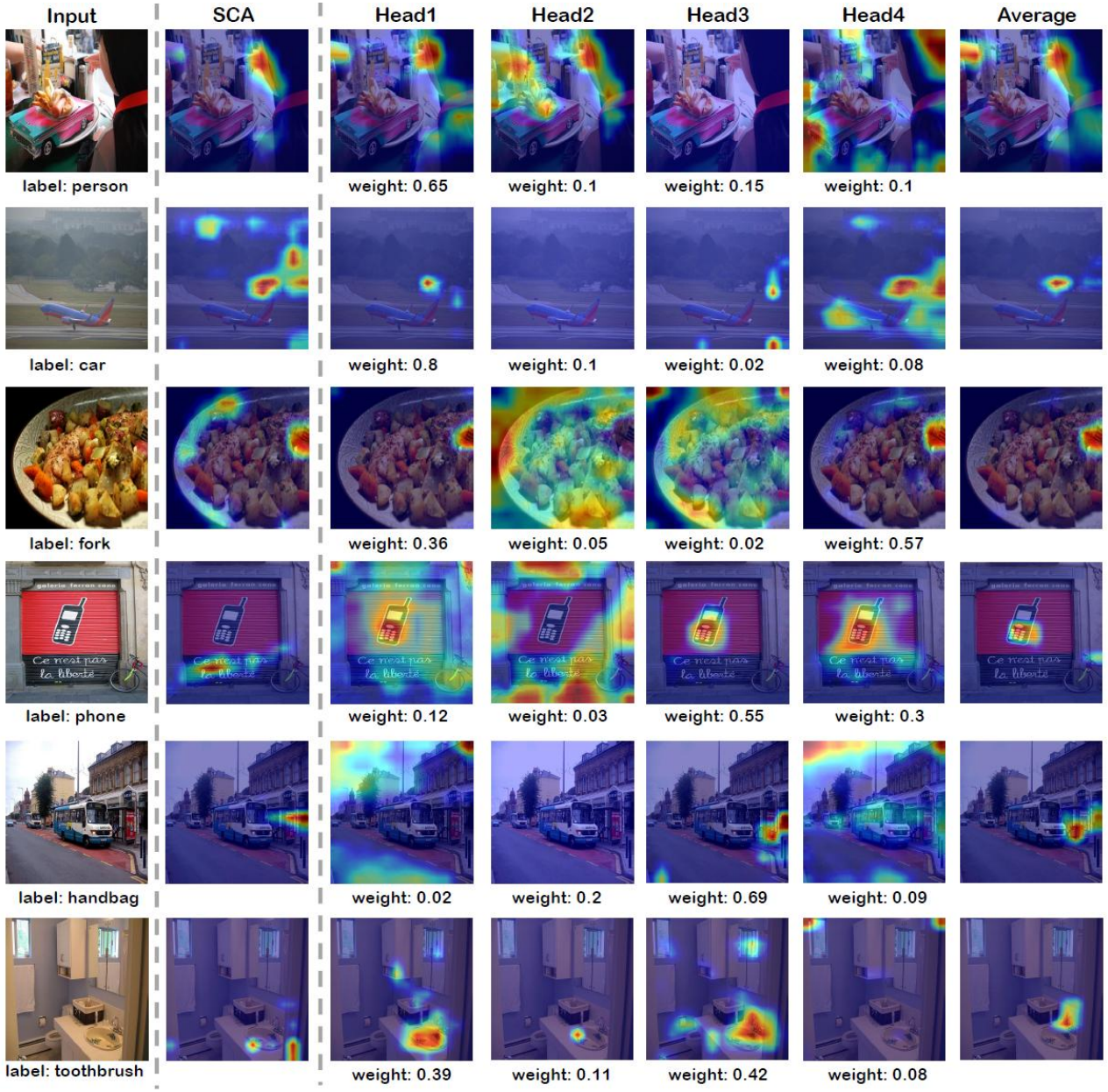
Method	OOD MS-COCO	Original MS-COCO
Baseline	50.0	81.4
CaaM	53.7	81.0
IDA-L	**58.6**	**84.3**

Experiments

Methods	Param.	FLOPs(448)	mAP(448)	mAP(576)
ResNet101	44.7M	31.4	81.5	82.7
CSRA (Zhu & Wu, 2021)	45.5M	31.7	83.5	-
C-Trans (Lanchantin et al., 2021)	45.0M	43.3	-	85.1
ADDGCN (Ye et al., 2020)	48.2M	32.6	-	85.2
CCD (Liu et al., 2022)	48.3M	32.0	84.0	85.3
TDRG (Zhao et al., 2021)	68.3M	73.7	84.6	86.0
IDA(L)	45.6M	31.7	84.3	85.5
IDA(H)	55.1M	33.8	84.8	86.3

Methods	mAP	mAP*
ResNet101	92.9	-
MLGCN(Chen et al., 2019c)	94.0	-
ASL(Ridnik et al., 2021)	94.6	95.8
CSRA(Zhu & Wu, 2021)	94.7	96.0
ADDGCN(Ye et al., 2020)	93.6	96.0
TDRG(Zhao et al., 2021)	**95.0**	-
IDA-R101(L)	94.5	96.1
IDA-R101(H)	**95.0**	**96.4**

Methods	mAP	mAP*
ResNet101	92.5	-
HCP(Wei et al., 2015)	90.5	-
RCP(Wang et al., 2016)	92.2	-
SSGRL(Chen et al., 2019b)	93.9	94.8
CSRA(Zhu & Wu, 2021)	94.1	95.2
ADDGCN(Ye et al., 2020)	-	95.5
IDA-R101(L)	94.6	95.9
IDA-R101(H)	**95.0**	**96.3**



Thanks