南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Visual Attention Consistency under Image Transforms for Multi-Label Image Classification

Hao Guo[‡], Kang Zheng[‡], Xiaochuan Fan[‡], Hongkai Yu[♯], Song Wang[†,‡,*]

[†]Tianjin University, [‡]University of South Carolina, [♯]University of Texas - Rio Grande Valley

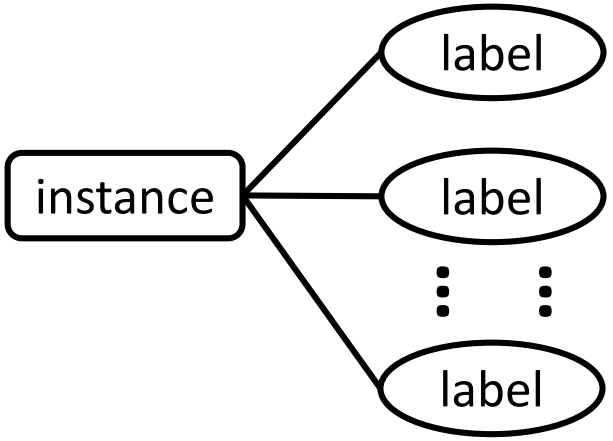
{hguo, zheng37}@email.sc.edu, efan3000@gmail.com, hongkai.yu@utrgv.edu, songwang@cec.sc.edu

CVPR 2019

Multi-Label Learning

- **Multi-label learning** vs. ordinary supervised learning

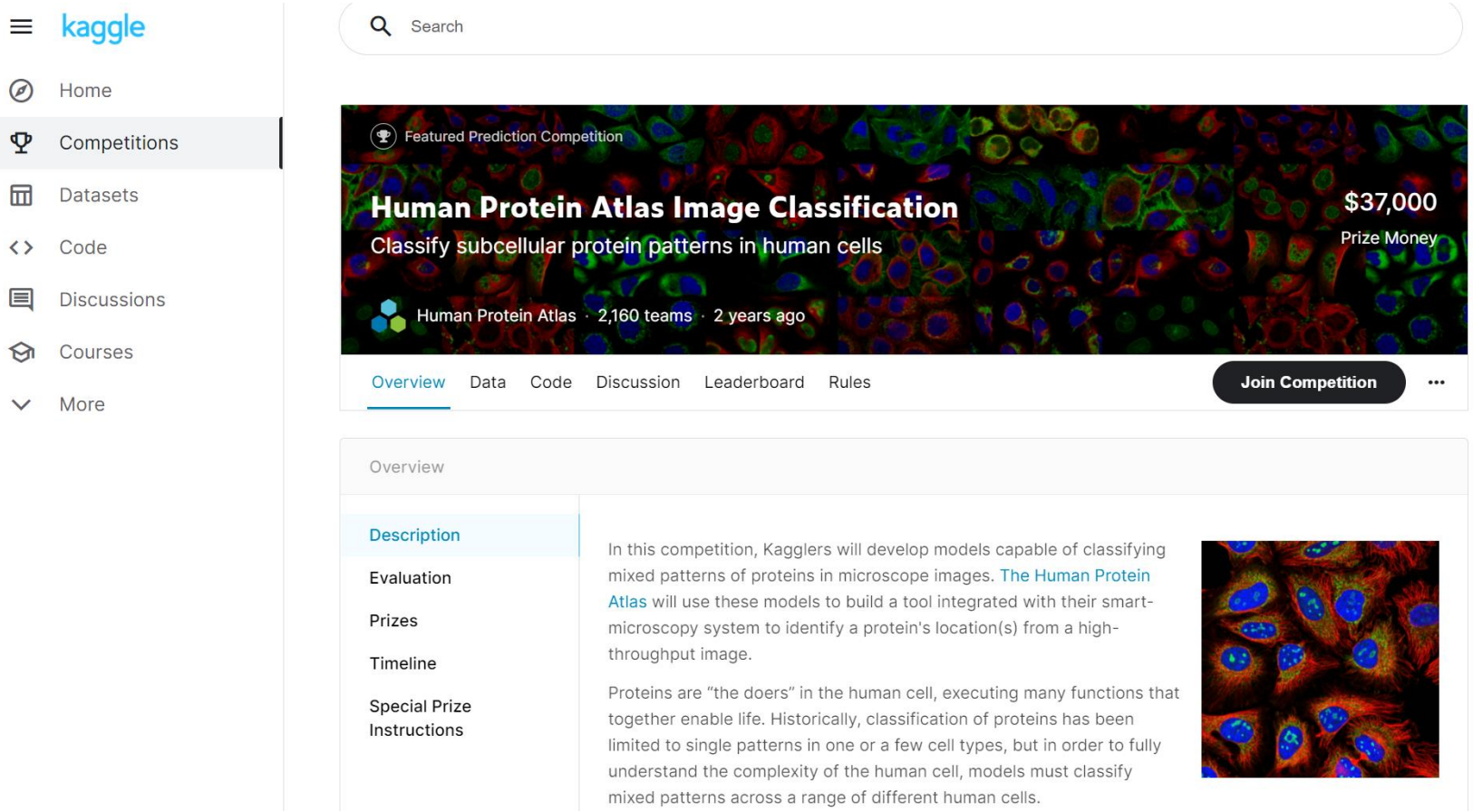
Ordinary supervised learning
(only one ground-truth label)



Multi-label Learning
(multiple ground-truth labels)

Multi Label Learning: Applications

- Human protein atlas image classification



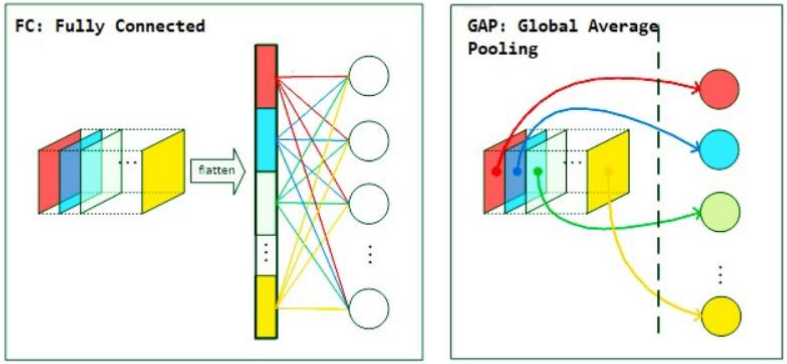
Multi-Label Learning: Applications

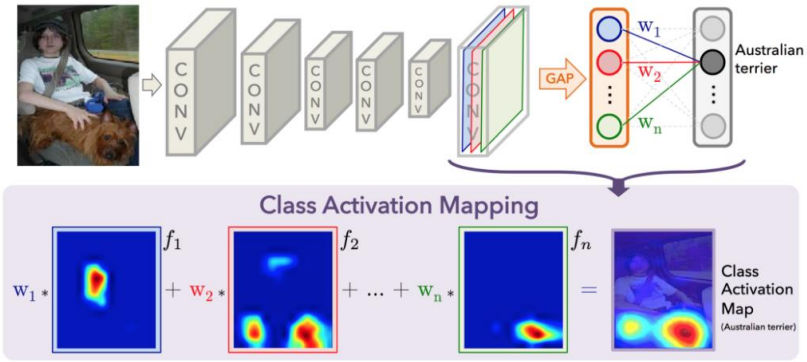
- Automatic Retail Checkout

[Wei et al., arxiv 2019]

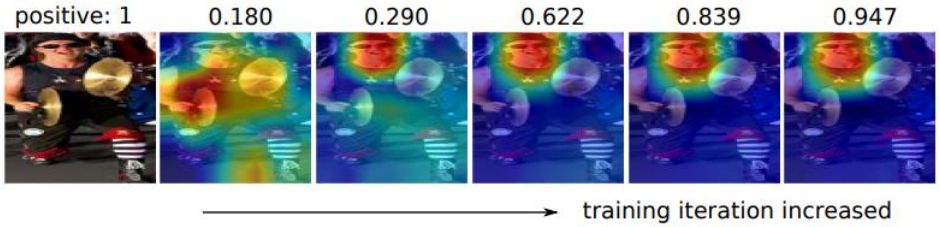
Background: Class Activation Mapping

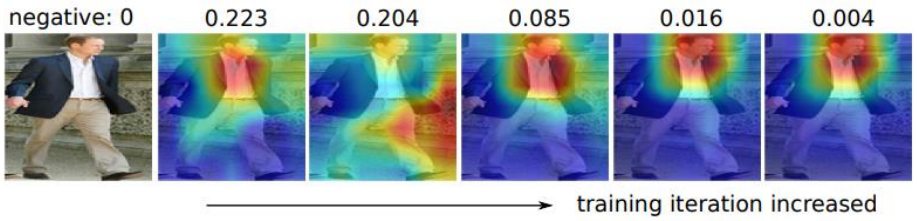
- Global Average Pooling (GAP)





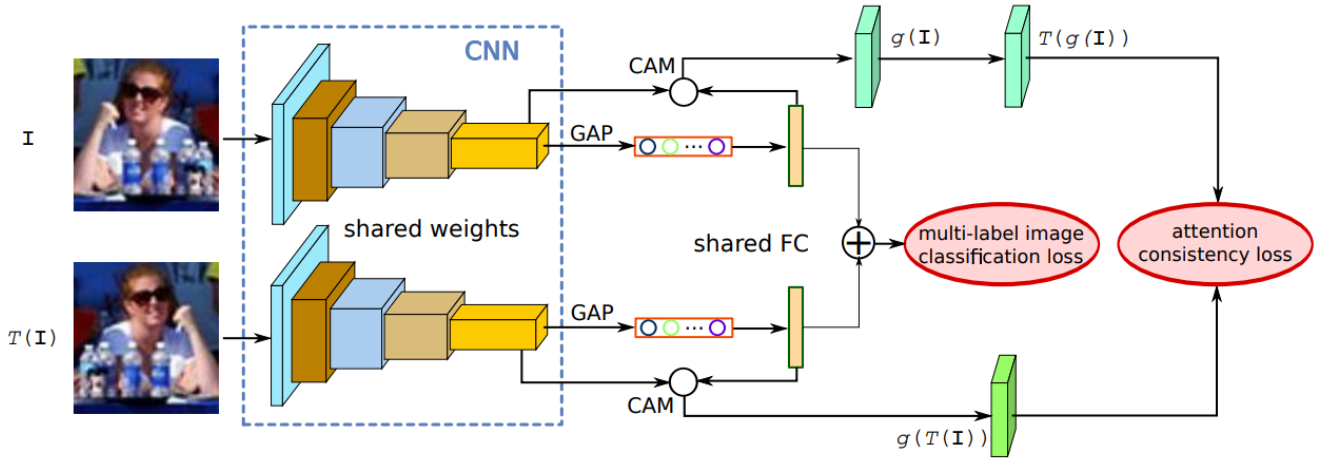
Attention heatmap $\boldsymbol{M} = g(\boldsymbol{I})$ $\boldsymbol{M}_j(m,n) = \sum_{k=1}^{C} \boldsymbol{W}(j,k)\boldsymbol{F}_k(m,n),$





The Proposed Network

- Two-branch network



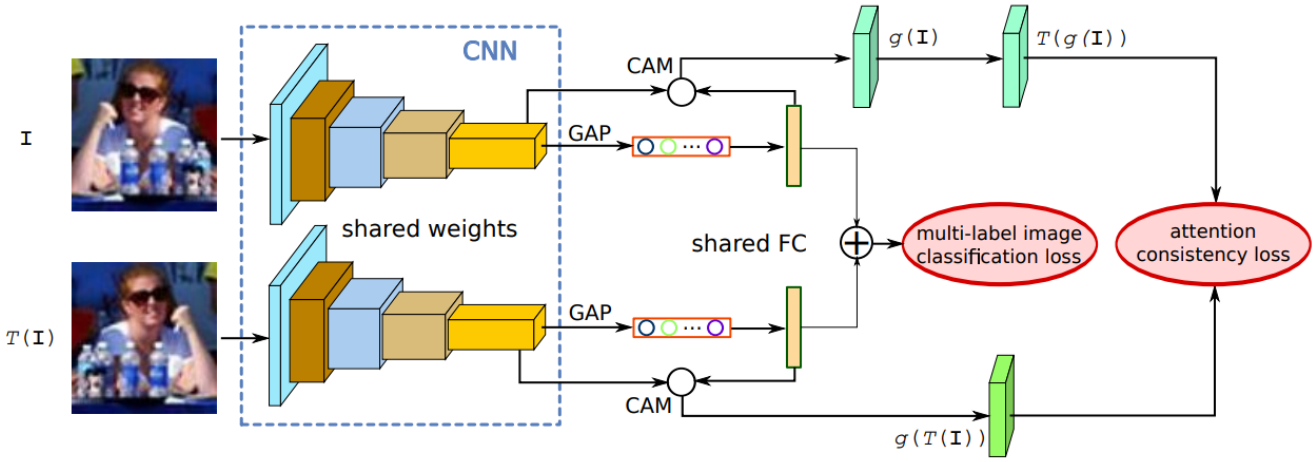
- Weighted binary cross entropy loss

$$\ell_c = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}\omega_{ij}\left(y_{ij}\log\frac{1}{1+e^{-x_{ij}}} + (1-y_{ij})\log\frac{e^{-x_{ij}}}{1+e^{-x_{ij}}}\right)$$

$$\omega_{ij} = \begin{cases} e^{1-p_j} & \text{if } y_{ij}=1 \\ e^{p_j} & \text{if } y_{ij}=0 \end{cases},$$

The Proposed Network

- Two-branch network



- Visual Attention Consistency

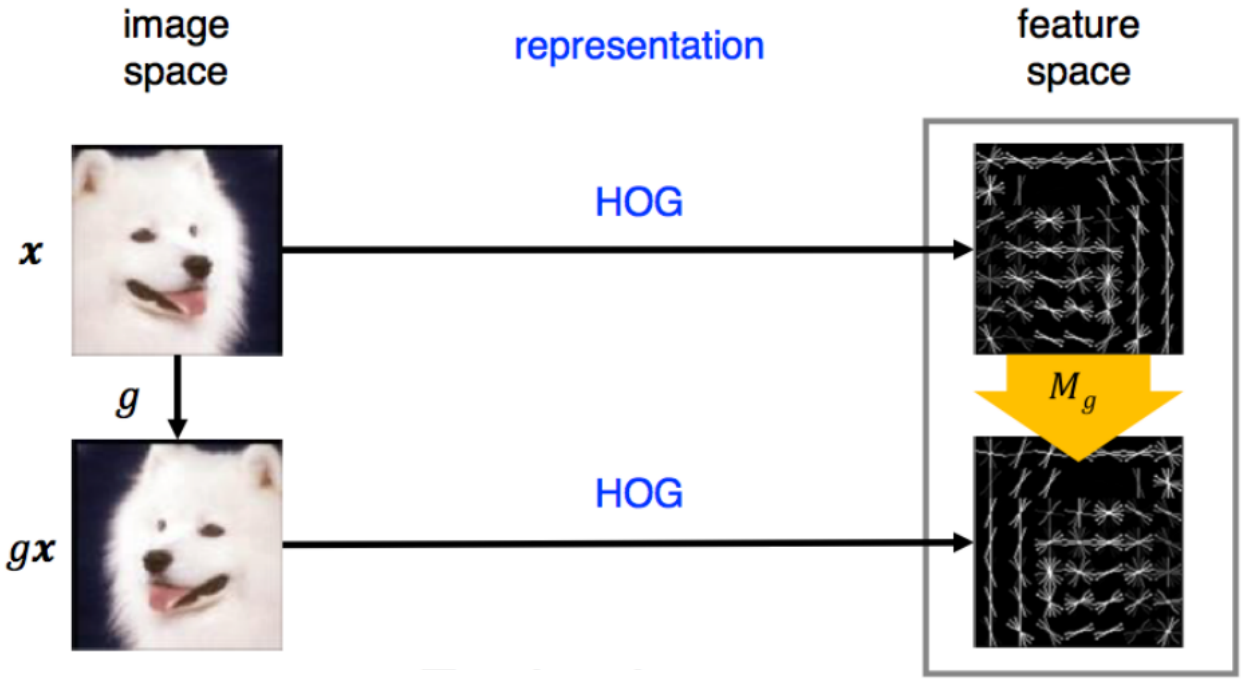
$$T(g(\boldsymbol{I})) = g(T(\boldsymbol{I})) \qquad \ell_a = \frac{1}{NLHW}\sum_{i=1}^{N}\sum_{j=1}^{L}\|\hat{\boldsymbol{M}}_{ij} - \boldsymbol{M'}_{ij}\|_2$$

- The joint objective function

$$\ell = \ell_c + \lambda\ell_a$$

Understanding Representation

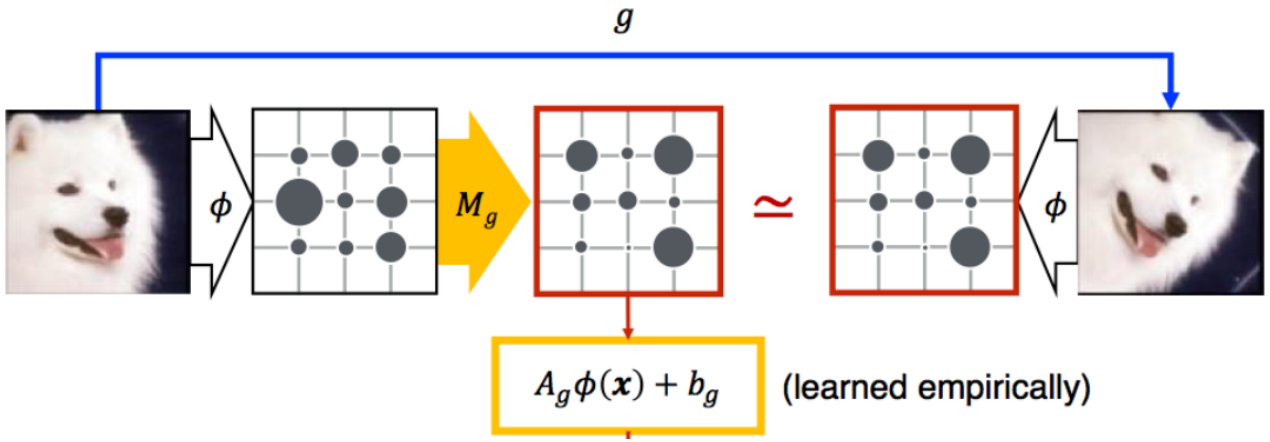
- Equivariance of Representation
  - Study how transformations of the input image are encoded by the representation



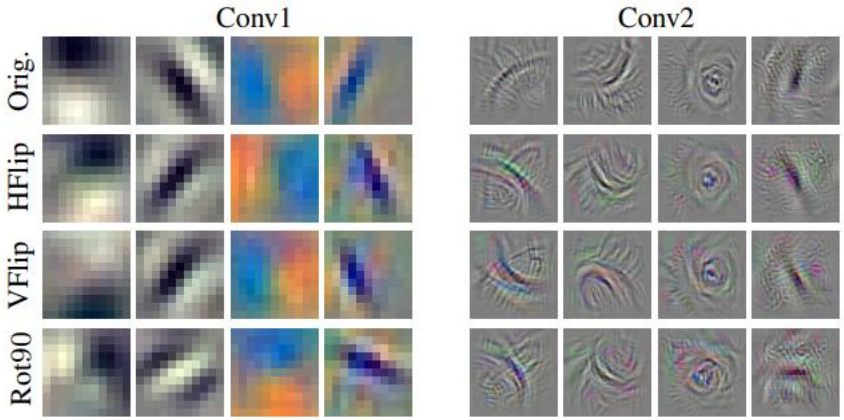
$$\phi(g\mathbf{x}) \approx M_g\phi(\mathbf{x})$$

Understanding Representation

- Learning the equivariance
  - Equivariant transformations can be learned empirically from data, and amount to simple linear transformation of the representation output



- Equivariant transformations:
  - Scaling, rotation, flipping, translation
- Equivariant representation:
  - HOG
  - Early convolutional layers in CNNs



Connection to Consistency Regularization

- Main differences

Impact to the network

$$\sum_{b=1}^{\mu B} \|p_{\mathrm{m}}(y|\,\alpha(u_b)) - p_{\mathrm{m}}(y|\,\alpha(u_b))\|_2^2 \quad < \quad \ell_a = \frac{1}{NLHW}\sum_{i=1}^{N}\sum_{j=1}^{L}\|\hat{\boldsymbol{M}}_{ij} - \boldsymbol{M'}_{ij}\|_2$$

- Different transformations
- Impose transforms on the final output (high-level representation).

- Single transformation
- The proposed method enforces attention consistency at middle-level representation.

Experiments

- Ablation studies

Table 1. Performance (%) on WIDER Attribute dataset in terms of label-based metrics. The best results are highlighted in bold font and red color, while the second bests are in blue.

model	mAP	mA	F1-C	P-C	R-C	F1-O	P-O	R-O
R50	83.4	82.0	73.9	79.5	69.4	79.4	82.3	76.6
R50+t	83.7	83.4	74.1	75.6	72.8	79.5	80.6	78.4
R50+r	83.2	82.8	73.2	75.9	71.1	78.5	81.0	76.1
R50+s	83.9	83.0	74.4	77.7	71.7	79.4	81.3	77.6
R50+f	84.2	82.8	74.6	79.5	70.7	80.0	82.9	76.9
R50+ACt	83.9	84.0	74.2	74.5	74.2	79.2	79.7	78.7
R50+ACr	85.0	83.3	75.1	79.2	71.8	80.2	82.3	77.9
R50+ACs	85.6	82.7	75.3	81.9	70.1	80.6	**84.5**	77.1
R50+ACf	86.3	**84.5**	76.4	78.9	**74.3**	81.2	82.6	**79.8**
R50+ACfs	**86.8**	83.7	**76.5**	**82.4**	72.1	**81.8**	84.4	79.3
R101	84.8	83.2	75.5	80.5	71.5	80.6	83.6	77.8
R101+ACt	84.6	83.5	75.3	79.1	71.9	80.1	83.1	77.3
R101+ACr	86.0	84.2	76.2	79.5	73.6	81.2	83.2	79.4
R101+ACs	86.5	83.6	76.5	**82.4**	71.9	81.6	**85.1**	78.3
R101+ACf	87.1	84.7	77.4	80.9	74.5	82.1	83.8	80.5
R101+ACfs	**87.5**	**85.0**	**77.6**	81.3	**74.8**	**82.4**	84.1	**80.7**

Experiments

- MS-COCO

Table 6. Performance (%) of the comparison methods and the proposed method on MS-COCO dataset with label-based metrics. The method ResNet101* represents the baseline used in work [67] implemented from the original ResNet101 [20] with complex data augmentations.

Method		All							top-3					
		mAP	F1-C	P-C	R-C	F1-O	P-O	R-O	F1-C	P-C	R-C	F1-O	P-O	R-O
WARP [17]		-	-	-	-	-	-	-	55.7	59.3	52.5	60.7	59.8	61.4
CNN-RNN [53]		-	-	-	-	-	-	-	60.4	66.0	55.6	67.8	69.2	**66.4**
ResNet101* [67]		75.2	69.5	80.8	63.4	74.4	82.1	68.0	65.9	84.3	57.4	71.7	86.5	61.3
ResNet101-SRN [67]		77.1	71.2	81.6	65.4	75.8	82.7	69.9	67.4	85.2	58.8	72.9	87.4	62.5
baseline	ResNet101	74.9	69.7	70.1	69.7	73.7	73.6	73.7	66.1	77.7	59.8	71.2	82.2	62.8
Ours	ResNet101-ACs	76.8	70.1	**83.3**	62.1	74.9	**85.7**	66.5	66.3	**87.6**	56.3	72.0	**89.6**	60.1
	ResNet101-ACf	77.3	71.9	73.5	**71.0**	75.7	76.5	**74.9**	67.9	81.9	**61.0**	73.0	84.5	64.2
	ResNet101-ACfs	**77.5**	**72.2**	77.4	68.3	**76.3**	79.8	73.1	**68.0**	85.2	59.4	**73.1**	86.6	63.3

Experiments

- WIDER

Table 4. Performance (%) of the comparison methods and the proposed method on WIDER in terms of label-based metrics. The method ResNet101* represents the baseline used in work [67] implemented from the original ResNet101 [20] with multiple data augmentations.

	method	mAP	F1-C	P-C	R-C	F1-O	P-O	R-O
	R-CNN [15]	80.0	-	-	-	-	-	-
	R*CNN [16]	80.5	-	-	-	-	-	-
	DHC [34]	81.3	-	-	-	-	-	-
	AR [18]	82.9	-	-	-	-	-	-
	ResNet101* [67]	85.0	74.7	-	-	80.4	-	-
	SRN [67]	86.2	75.9	-	-	81.3	-	-
	VAA [44]	86.4	-	-	-	-	-	-
Ours	R50	83.4	73.9	79.5	69.4	79.4	82.3	76.6
Ours	R50+ACs	85.6	75.3	81.9	70.1	80.6	84.5	77.1
Ours	R50+ACf	86.3	76.4	78.9	74.3	81.2	82.6	79.8
Ours	R50+ACfs	86.8	76.5	**82.4**	72.1	81.8	84.4	79.3
Ours	R101	84.8	75.5	80.5	71.5	80.6	83.6	77.8
Ours	R101+ACs	86.5	76.5	**82.4**	71.9	81.6	**85.1**	78.3
Ours	R101+ACf	87.1	77.3	80.9	74.5	82.1	83.8	80.5
Ours	R101+ACfs	**87.5**	**77.6**	81.3	**74.8**	**82.4**	84.1	**80.7**

Experiments

- Visualization

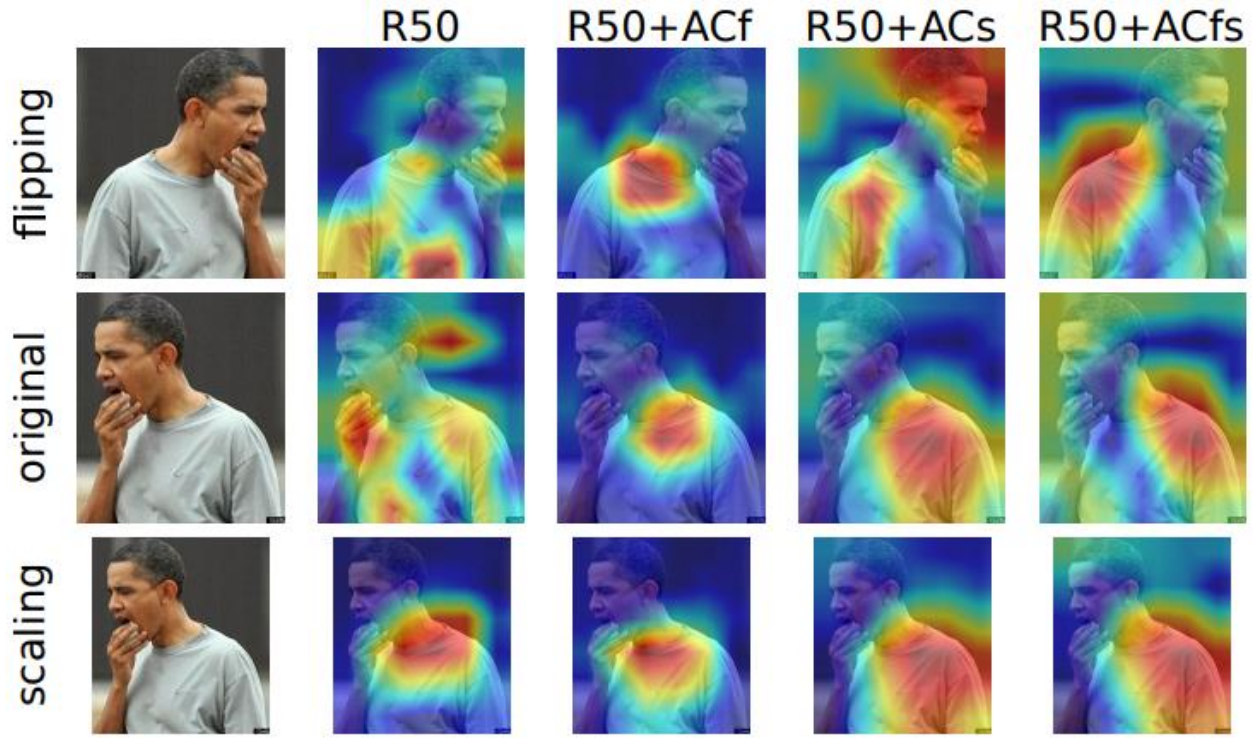


Figure 5. Attention heatmaps for classifying label "T-shirt" from flipped (row 1), original (row 2), and scaled (row 3) images using different models.

THANKS