Paper Reading

蒋 磊

2018.12.1

Outline

- **Learning Deep Features for Discriminative Localization, CVPR 2016**
- **Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, ICCV 2017**
- **Tell Me Where to Look: Guided Attention Inference Network, CVPR 2018**

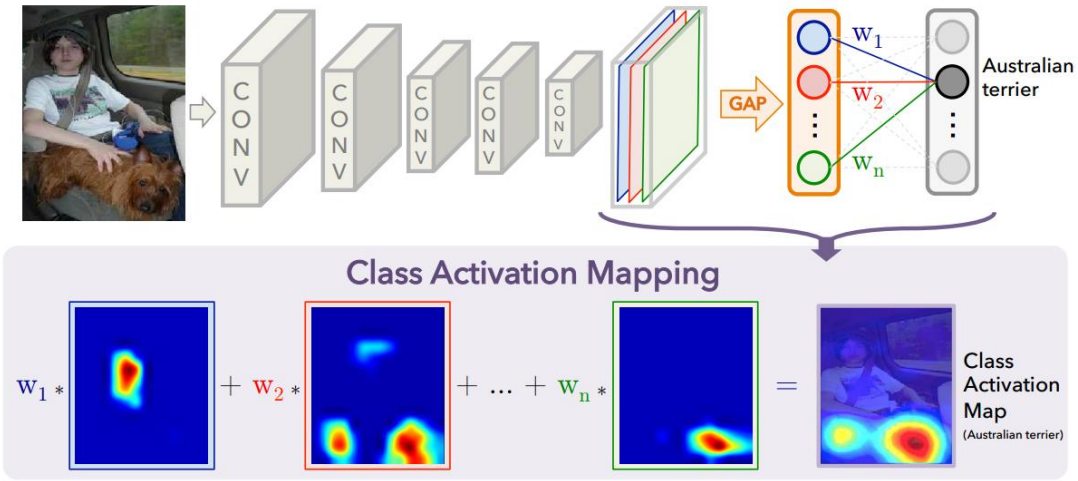
- Motivation

CNN have remarkable localization ability despite being trained on image level labels.

- Application

Weakly supervised object localization and CNN visualization

Object detectors emerge in deep scene cnns, ICLR 2015



$$F_k = \sum_{x,y} f_k(x,y)$$

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \sum_k w_k^c f_k(x,y).$$

$$M_c(x,y) = \sum_k w_k^c f_k(x,y).$$

$$S_c = \sum_{x,y} M_c(x,y)$$

Mc(x, y) directly indicates the importance of the activation at spatial grid(x, y) leading to the classification of an image to class c.

Learning Deep Features for Discriminative Localization , CVPR 2016

Results

Table 1. Classification error on the ILSVRC validation set.

Networks	top-1 val. error	top-5 val. error
VGGnet-GAP	33.4	12.2
GoogLeNet-GAP	35.0	13.2
AlexNet*-GAP	44.9	20.9
AlexNet-GAP	51.1	26.3
GoogLeNet	31.9	11.3
VGGnet	31.2	11.4
AlexNet	42.6	19.5
NIN	41.9	19.6
GoogLeNet-GMP	35.6	13.9

Table 2. Localization error on the ILSVRC validation set. *Backprop* refers to using [23] for localization instead of CAM.

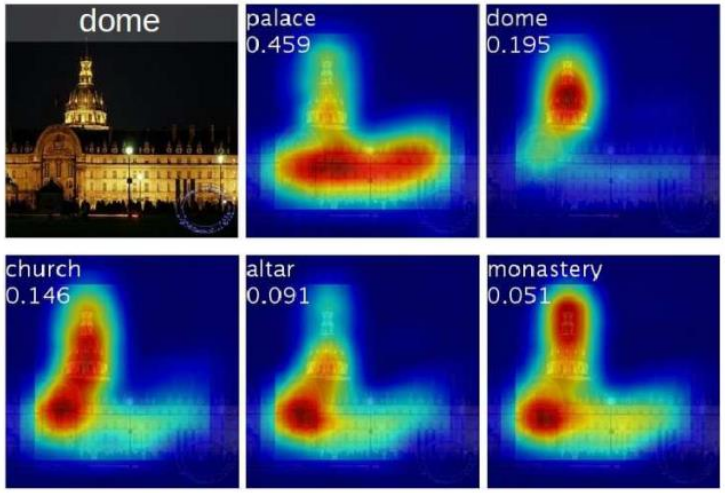
Method	top-1 val.error	top-5 val. error
GoogLeNet-GAP	**56.40**	**43.00**
VGGnet-GAP	57.20	45.14
GoogLeNet	60.09	49.34
AlexNet*-GAP	63.75	49.53
AlexNet-GAP	67.19	52.16
NIN	65.47	54.19
Backprop on GoogLeNet	61.31	50.55
Backprop on VGGnet	61.12	51.46
Backprop on AlexNet	65.17	52.64
GoogLeNet-GMP	57.78	45.26

Table 3. Localization error on the ILSVRC test set for various weakly- and fully- supervised methods.

Method	supervision	top-5 test error
GoogLeNet-GAP (heuristics)	weakly	**37.1**
GoogLeNet-GAP	weakly	42.9
Backprop [23]	weakly	46.4
GoogLeNet [25]	full	26.7
OverFeat [22]	full	29.9
AlexNet [25]	full	34.2

We first segment the regions of which the value is above 20% of the max value of the CAM. Then we take the bounding box that covers the largest connected component in the segmentation map.

We follow a slightly different bounding box selection strategy here: we select two bounding boxes (one tight and one loose) from the class activation map of the top 1st and 2nd predicted classes and one loose bounding boxes from the top 3rd predicted class. This heuristic is a trade-off between classification accuracy and localization accuracy.



Dome: 圆屋顶

Palace：宫廷，豪宅

church:：教堂

Altar：祭坛，圣坛

Monastery：修道院

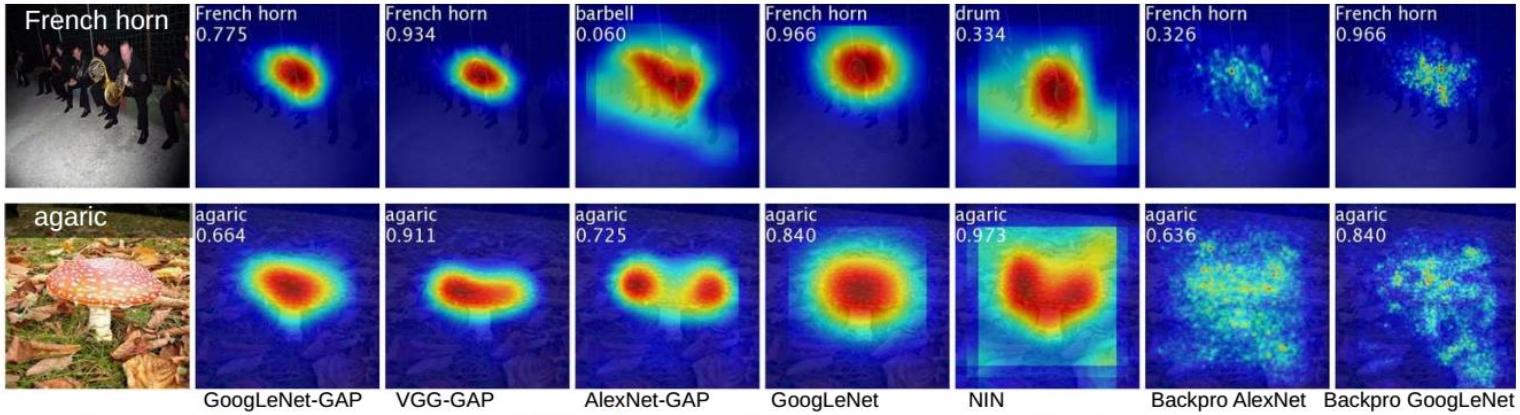
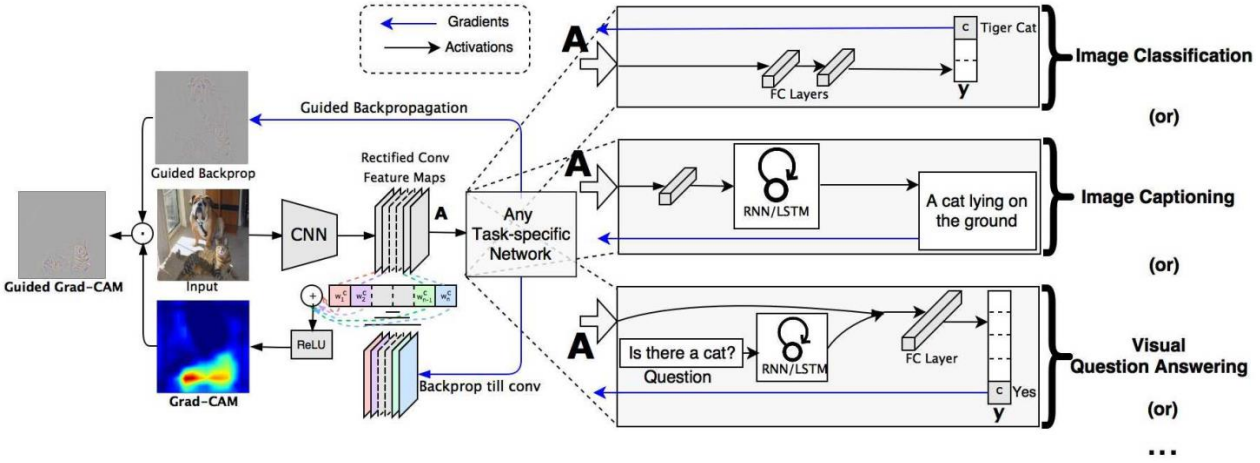


Figure 5. Class activation maps from CNN-GAPs and the class-specific saliency map from the backpropagation methods.



$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$S^c = \sum_k \underbrace{w_k^c}_{\text{class feature weights}} \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{A_{ij}^k}_{\text{feature map}}$$

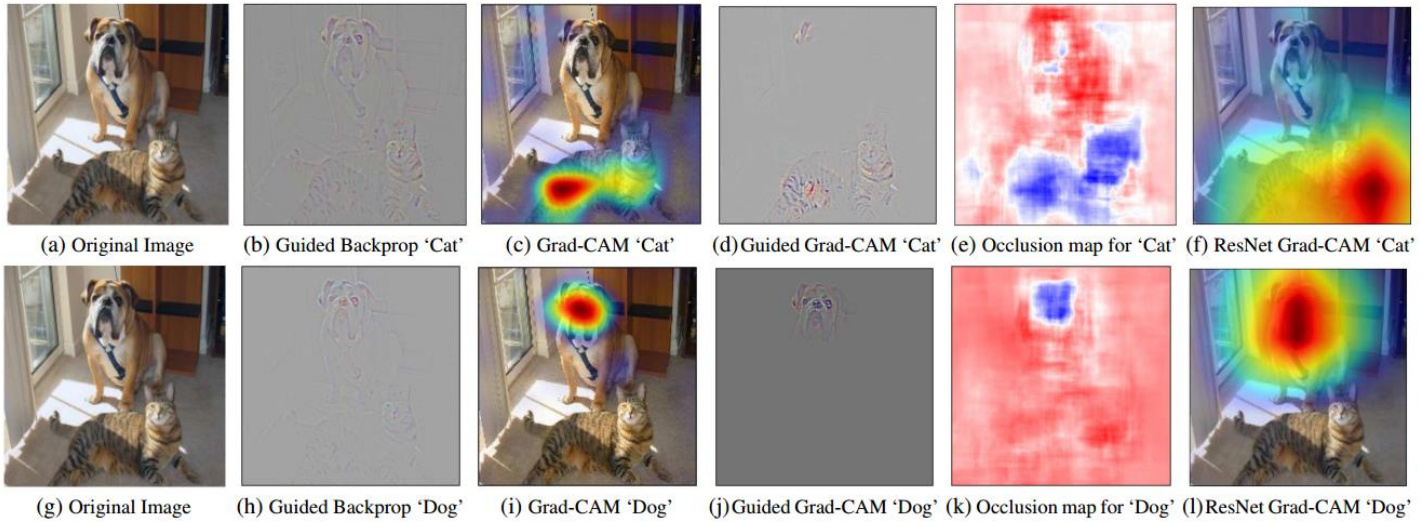
Grad-CAM as a generalization to CAM

$$L_{\text{Grad-CAM}}^c = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

$$S^c = \frac{1}{Z}\sum_i\sum_j \underbrace{\sum_k w_k^c A_{ij}^k}_{L_{\text{CAM}}^c}$$

We apply a ReLU to the linear combination of maps because we are only interested in the features that have a *positive* influence on the class of interest, *i.e.* pixels whose intensity should be *increased* in order to increase yc. Negative pixels are likely to belong to other categories in the image.

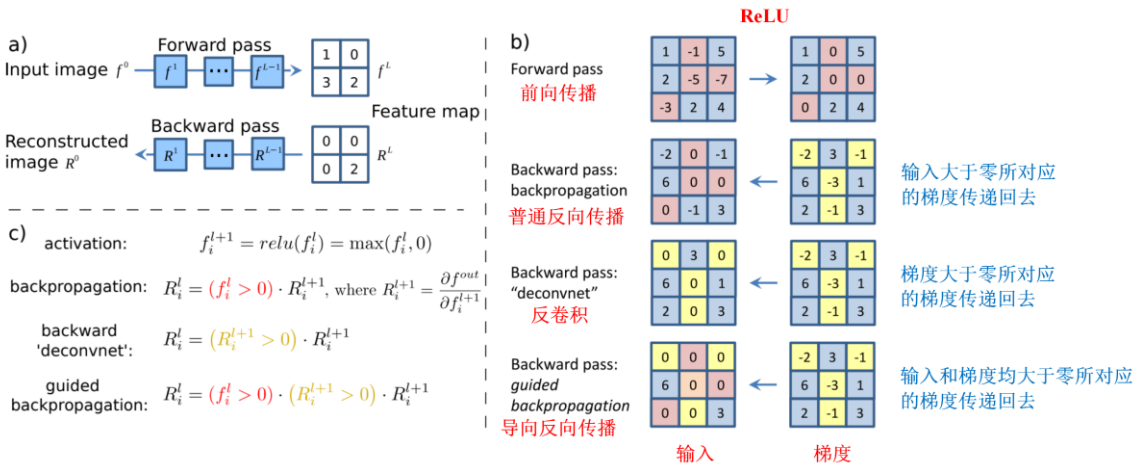
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, ICCV 2017



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d) Guided Grad-CAM 'Cat' (e) Occlusion map for 'Cat' (f) ResNet Grad-CAM 'Cat'

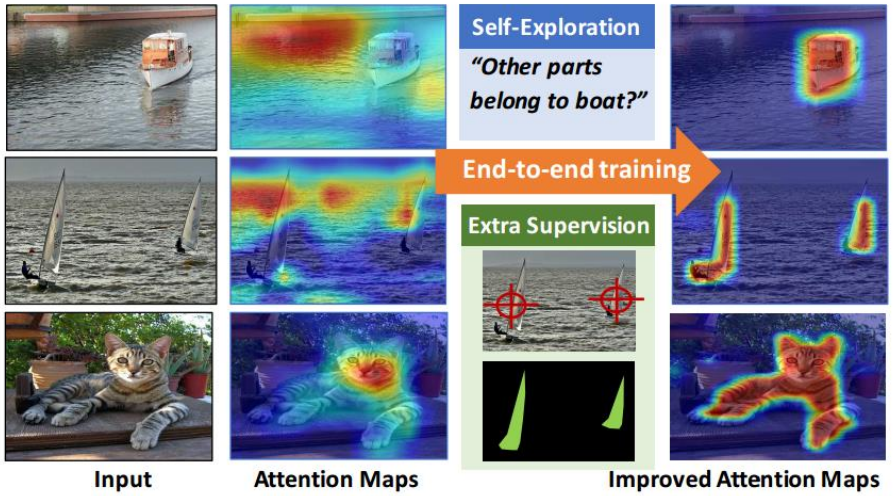
(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j) Guided Grad-CAM 'Dog' (k) Occlusion map for 'Dog' (l) ResNet Grad-CAM 'Dog'

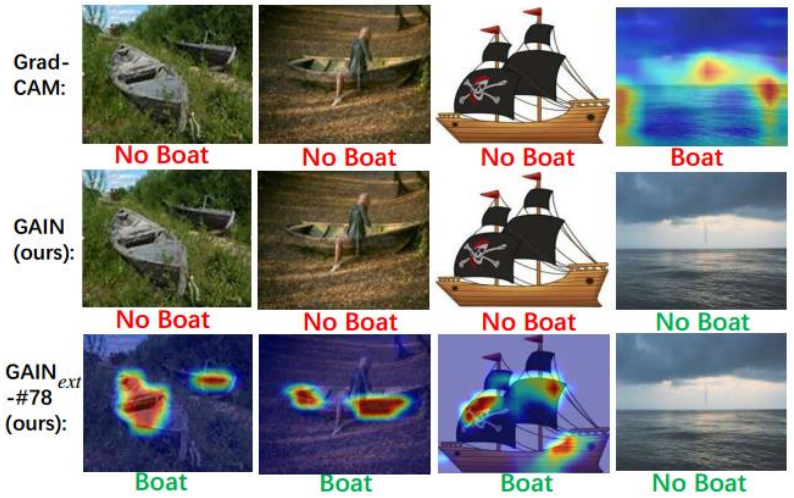
Figure 1: (a) Original image with a cat and a dog. (b-f) Support for the cat category according to various visualizations for VGG-16 and ResNet. (b) Guided Backpropagation [42]: highlights all contributing features. (c, f) Grad-CAM (Ours): localizes class-discriminative regions, (d) Combining (b) and (c) gives Guided Grad-CAM, which gives high-resolution class-discriminative visualizations. Interestingly, the localizations achieved by our Grad-CAM technique, (c) are very similar to results from occlusion sensitivity (e), while being orders of magnitude cheaper to compute. (f, l) are Grad-CAM visualizations for ResNet-18 layer. Note that in (c, f, i, l), red regions corresponds to high score for class, while in (e, k), blue corresponds to evidence for the class. Figure best viewed in color.



Method	Top-1 loc error	Top-5 loc error	Top-1 cls error	Top-5 cls error
Backprop on VGG-16 [40]	61.12	51.46	30.38	10.89
c-MWP on VGG-16 [46]	70.92	63.04	30.38	10.89
Grad-CAM on VGG-16 (ours)	56.51	46.41	30.38	10.89
VGG-16-GAP (CAM) [47]	57.20	45.14	33.40	12.20

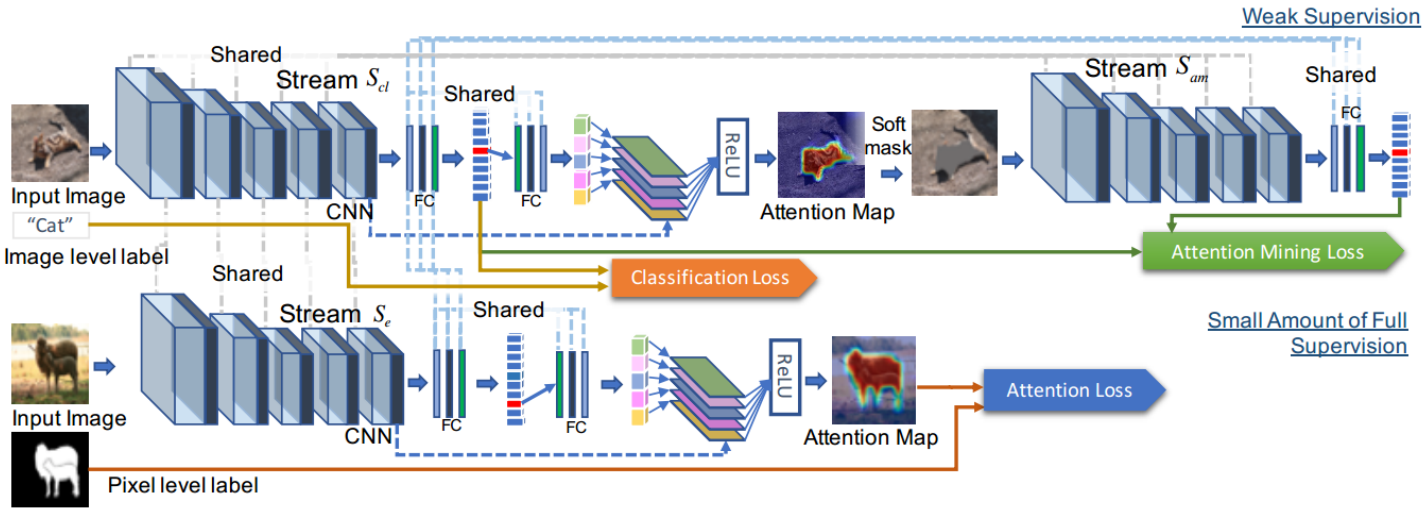
Table 1: Classification and Localization on ILSVRC-15 val (lower is better).





- Supervised by only classification loss, attention maps often only cover small and most discriminative regions of object of interest
- Bias in the training data(the foreground object incidentally always correlates with the same background object)

Tell Me Where to Look: Guided Attention Inference Network , CVPR 2018



$$T(A^c) = \frac{1}{1 + \exp(-\omega(A^c - \boldsymbol{\sigma}))}$$

$$I^{*c} = I - (T(A^c) \odot I),$$

$$L_{am} = \frac{1}{n} \sum_c s^c(I^{*c}),$$

$$L_e = \frac{1}{n} \sum_c (A^c - H^c)^2,$$

$$L_{ext} = L_{cl} + \alpha L_{am} + \omega L_e,$$

where $s^c(I^{*c})$ denotes the prediction score of I^{*c} for class c. n is the number of ground-truth class labels for this image I.

Results

Methods	Training Set	*val.* (mIoU)	*test* (mIoU)
Supervision: Purely Image-level Labels			
CCNN [19]	10K weak	35.3	35.6
MIL-sppxl [20]	700K weak	35.8	36.6
EM-Adapt [18]	10K weak	38.2	39.6
DCSM [25]	10K weak	44.1	45.1
BFBP [23]	10K weak	46.6	48.0
STC [32]	50K weak	49.8	51.2
AF-SS [21]	10K weak	52.6	52.7
CBTS-cues [22]	10K weak	52.8	53.7
TPL [11]	10K weak	53.1	53.8
AE-PSL [31]	10K weak	55.0	55.7
SEC [12] (baseline)	10K weak	50.7	51.7
GAIN (ours)	10K weak	**55.3**	**56.8**
Supervision: Image-level Labels (* Implicitly use pixel-level supervision)			
MIL-seg* [20]	700K weak + 1464 pixel	40.6	42.0
TransferNet* [9]	27K weak + 17K pixel	51.2	52.1
AF-MCG* [21]	10K weak + 1464 pixel	54.3	55.5
GAIN$_{ext}$* (ours)	10K weak + 200 pixel	**58.3**	**59.6**
GAIN$_{ext}$* (ours)	10K weak + 1464 pixel	**60.5**	**62.1**

Table 1. Comparison of weakly supervised semantic segmentation methods on PASCAL VOC 2012 *segmentation val.* set and *segmentation test* set. **weak** denotes image-level labels and **pixel** denotes pixel-level labels. *Implicitly use pixel-level supervision* is a protocol we followed as defined in [31], that pixel-level labels are only used in training priors, and only weak labels are used in the training of segmentation framework, e.g. SEC [12] in our case.

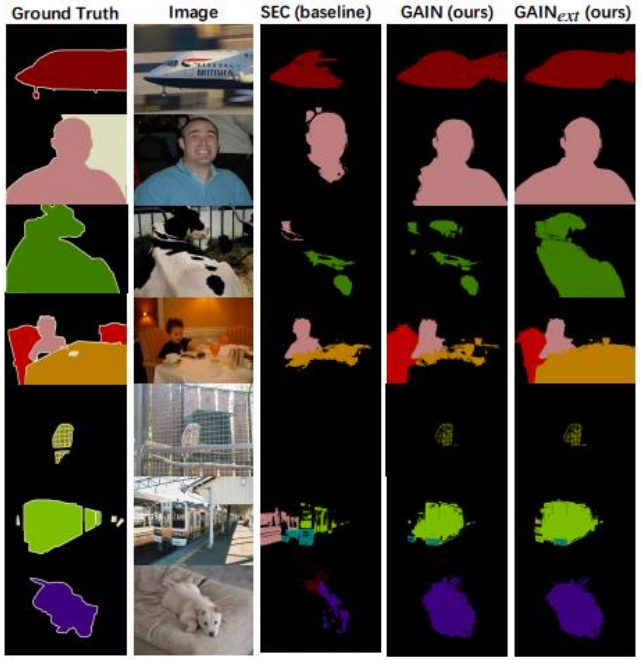


Figure 4. Qualitative results on PASCAL VOC 2012 *segmentation val.* set. They are generated by SEC (our baseline framework), our GAIN-based SEC and GAIN$_{ext}$-based SEC implicitly using 200 randomly selected (2%) extra supervision.

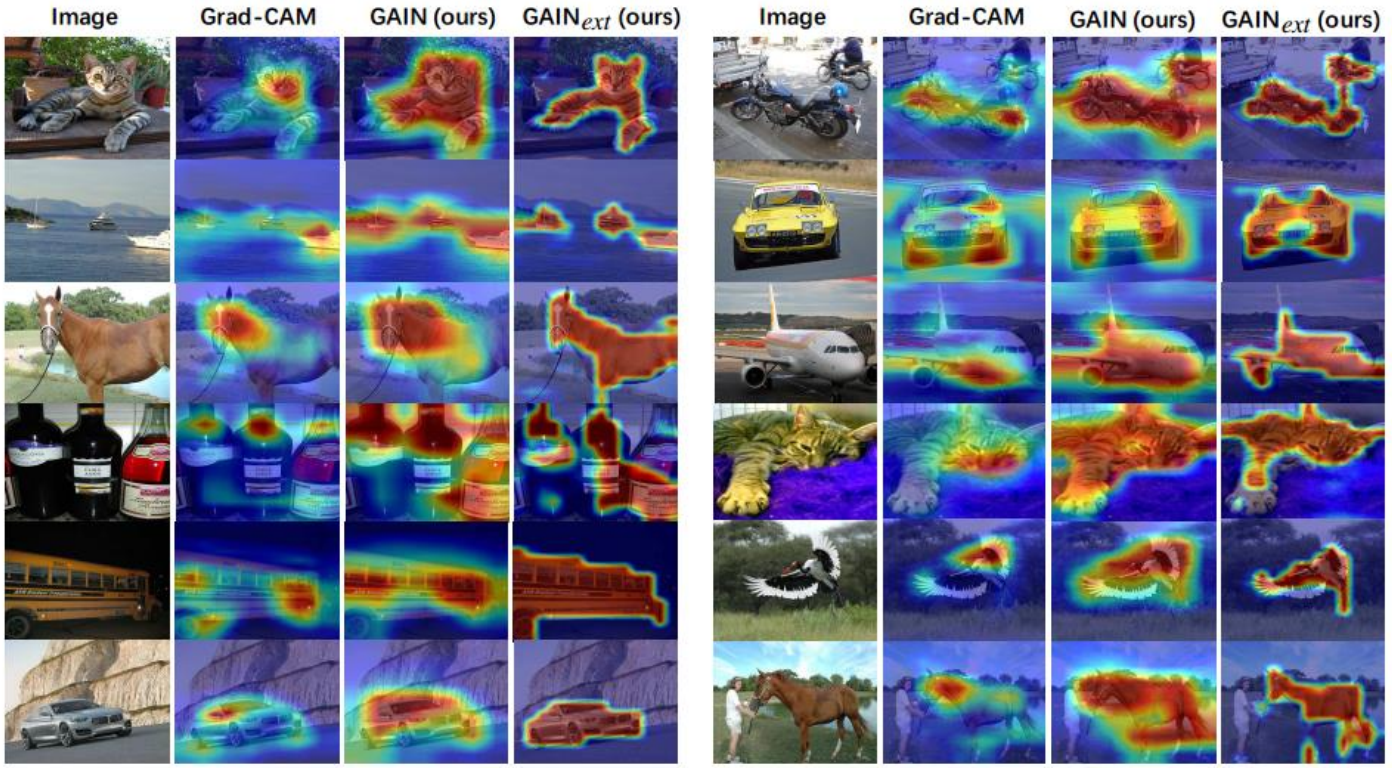


Figure 5. Qualitative results of attention maps generated by Grad-CAM [24], our GAIN and GAIN$_{ext}$ using 200 randomly selected (2%) extra supervision.

THE END!