Weakly-supervised object localization via class activation mapping

José Oramas

Internet Data Lab (IDLab), University of Antwerp, imec.

Personal Context

Some details

Teaching
- Operating Systems (1500WETOPS)
- Distributed Systems (1500WETDIS)
- Artificial Neural Networks (2500WETANN)

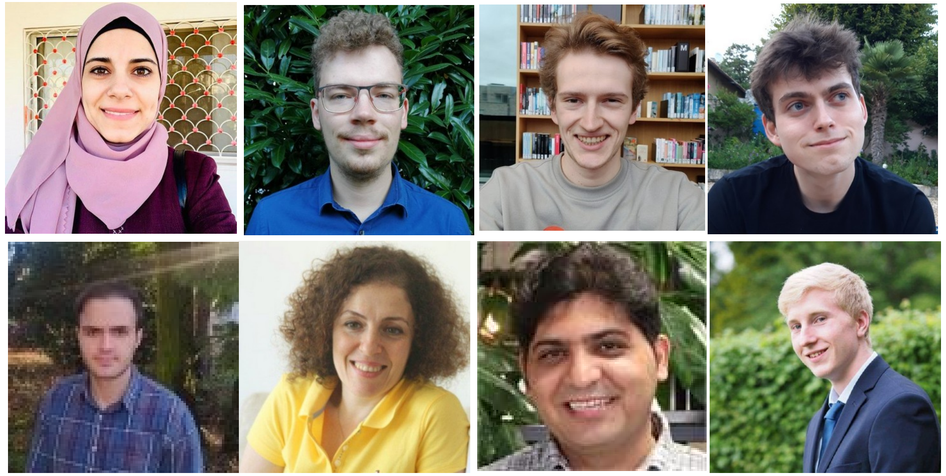
Affiliations
- Internet Data Lab (IDLab)

Research Interests
- Representation Learning
- Computer Vision
- Explainable AI / Interpretable ML

Research Lab

Internet Data Lab (IDLab) – UAntwerp, imec

Research Team



First of all

Research is Team Sport

Kaili Wang

José Oramas

Tinne Tuytelaars

Background

Background

Computer Vision – In Theory

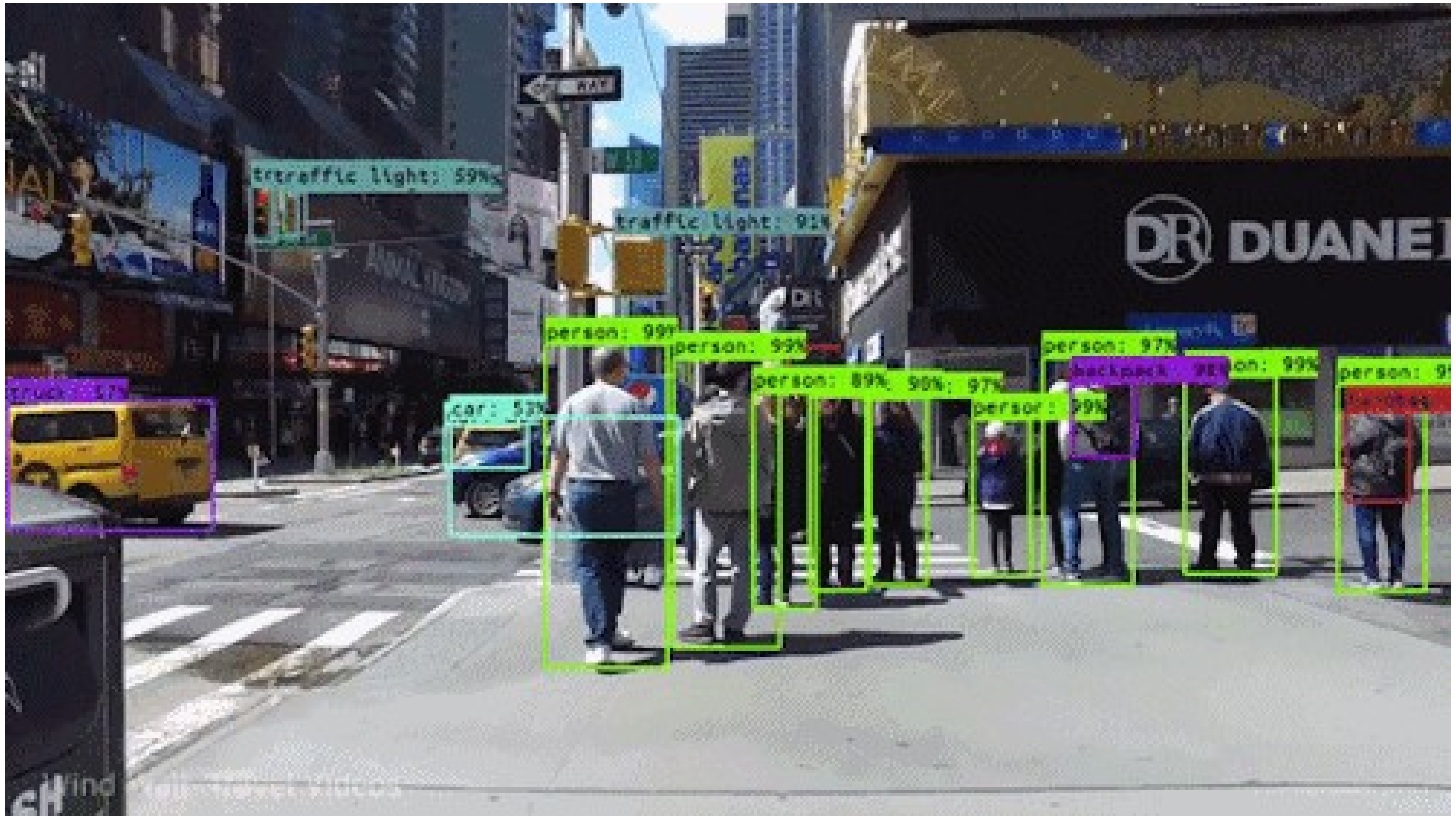
- **Objective:**

  Provide Computer Systems with the Sense of Sight we Possess

Background

Computer Vision – In Practice

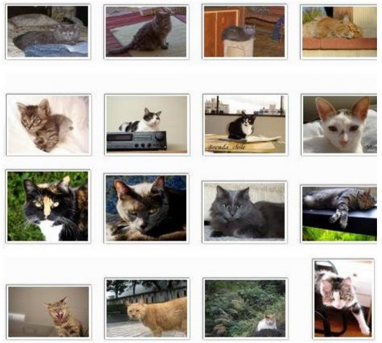
- Recognize and localize objects, actions, etc. in visual data (images and videos)



Background

Computer Vision – Classical Approach

dataset



Feature extraction

Interest Points
(SIFT, SURF, kAS, etc.)

Machine Learning

ML Method
(SVM, RF, Boosting, etc.)

Prediction

- **Idea:** Engineer informative features + Use ML to discriminate between those features

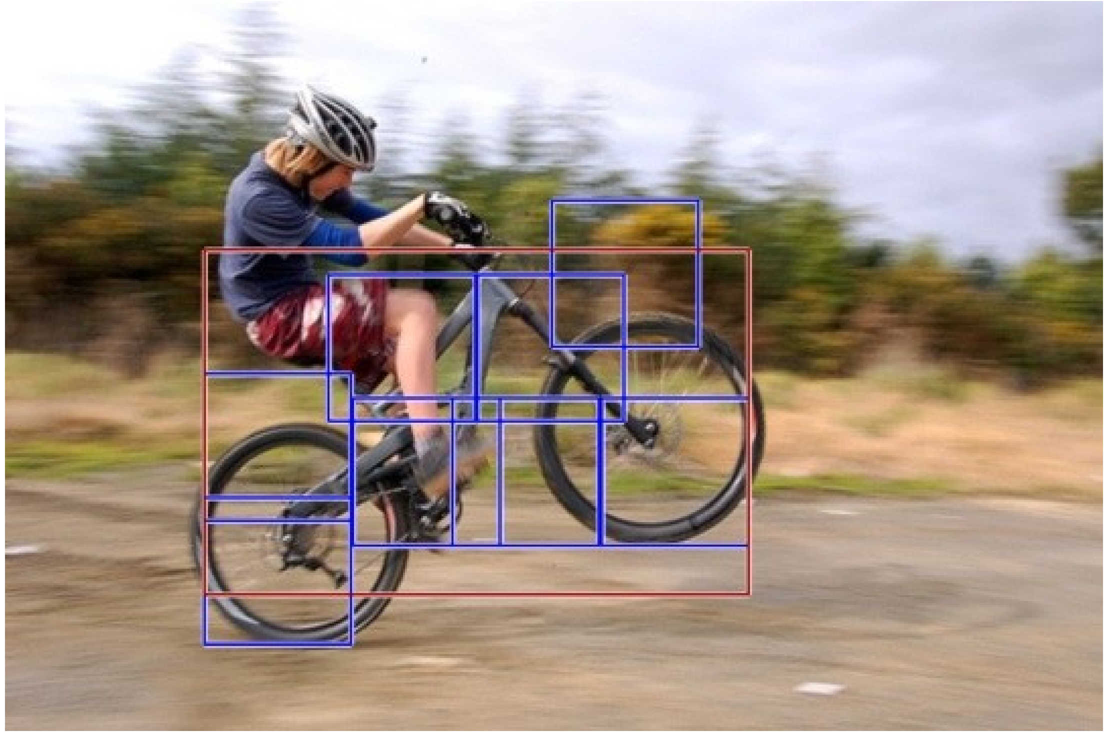
Background

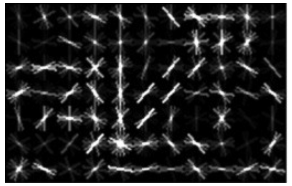
Computer Vision – Classical Approach

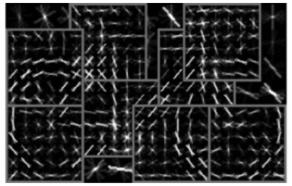
- **How to do that?**

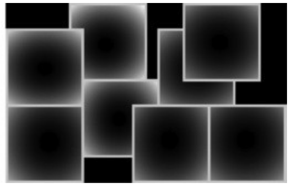
  Deformable-Part Models (DPMs)

  [Felzenszwalb et al., TPAMI'10]







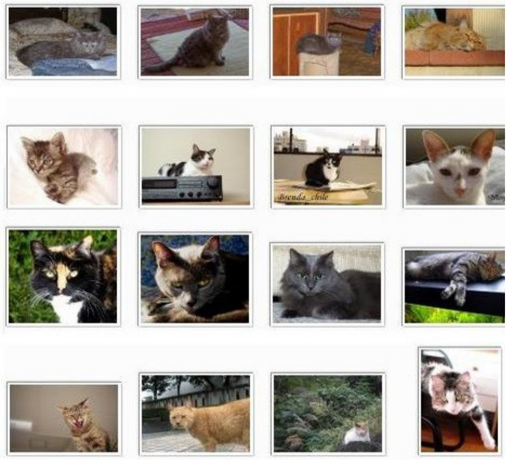


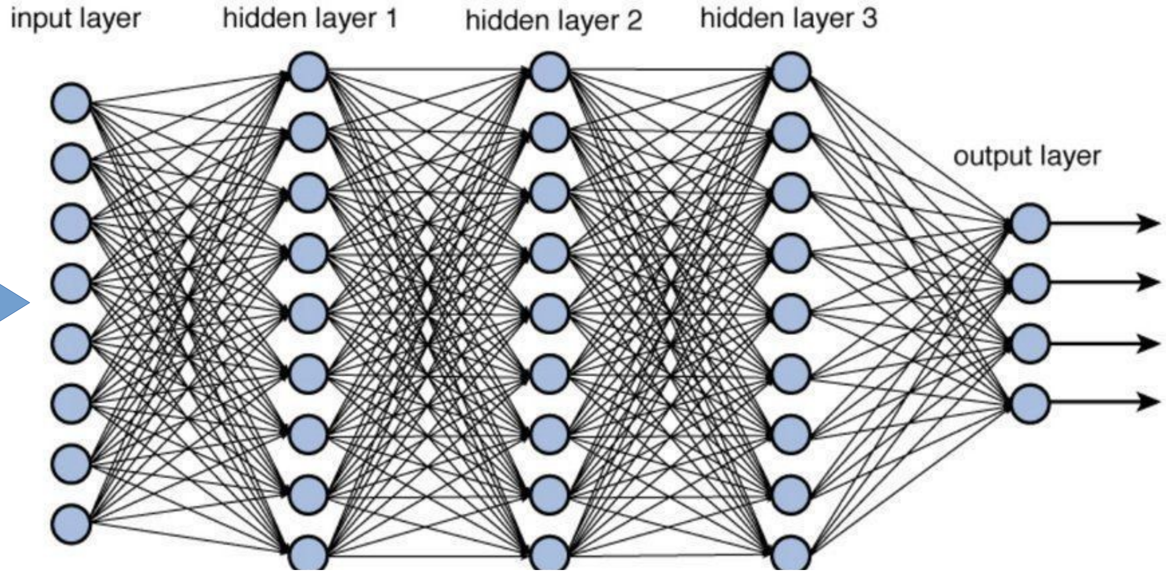
Background

Learning-based Representations

Deep Neural Network

Dataset





"Siamese cat"

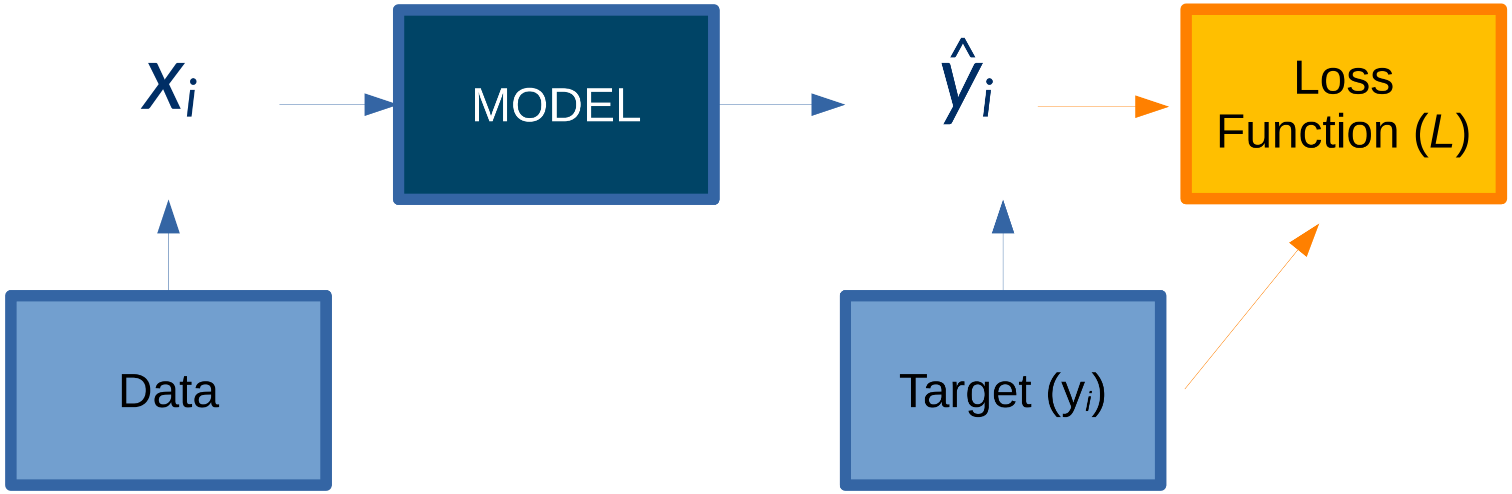
- **Idea:** Let the ML method figure out what features are important

  (i.e. Representation Learning)

Training a Model

Given:

- *Classification Task* with *k* classes.
- Training Data: inputs (x_i) and labels (y_i)



Object Localization

[… Reducing the Level of Required Annotation]

Background: Object Detection

Given: an input image X_i

Do: predict a label c_i *(out of a set of class labels)* & *location* *(bounding box)*

Required Data

– X_i

– $Y_i = \{ c_i, x_i, y_i, h_i, w_i \}$

Task of Interest: Object Localization

Given: an input image X_i *and a* prediction a label c_i *(out of a set of class labels).*

Do: *predict the location (bounding box)*

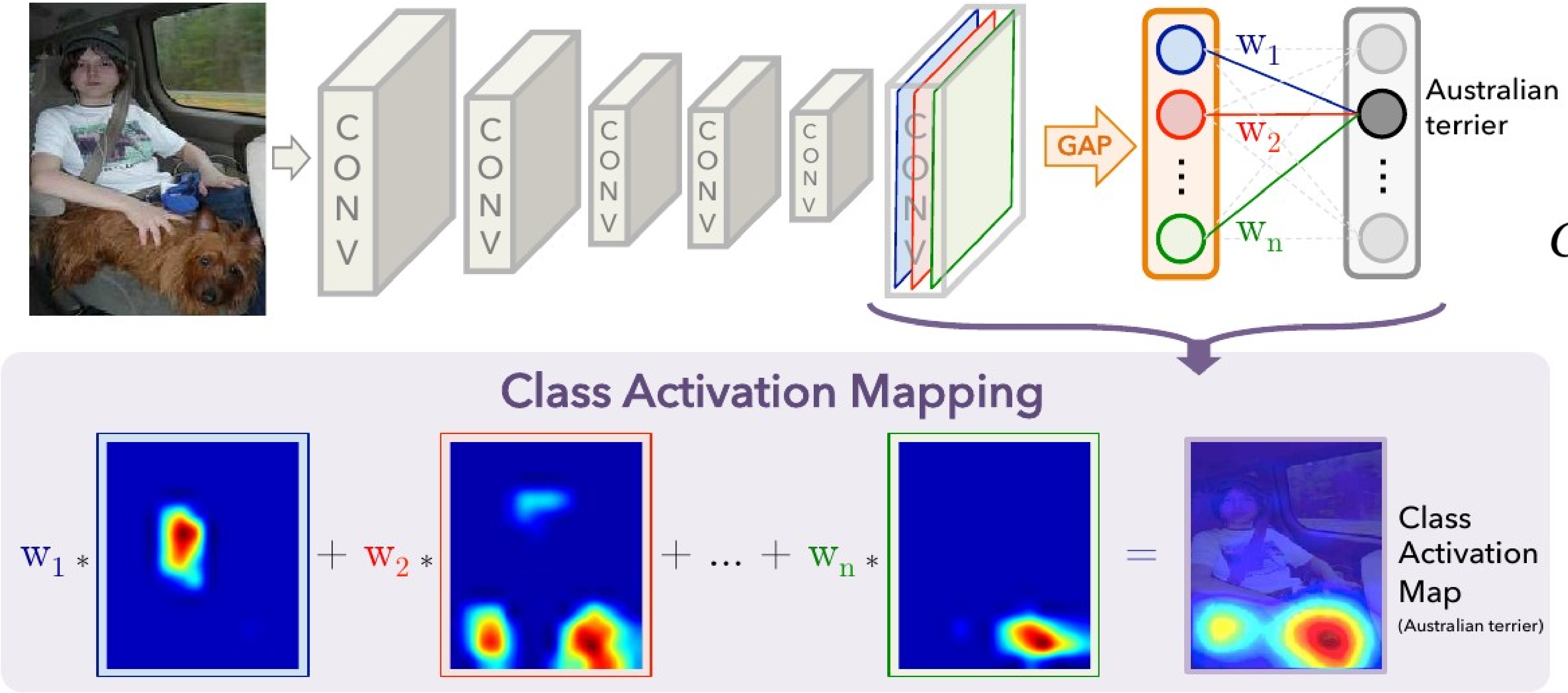
Required Data

– X_i

– $Y_i = \{ c_i \}$

Weakly-supervised Localization via CAM

Class Activation Mapping (CAM) [Zhou et al., 2016]

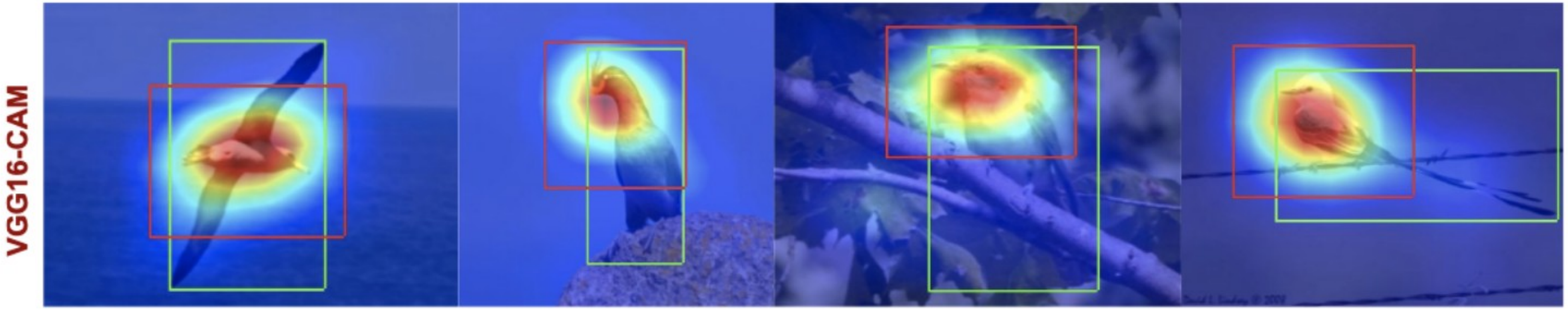


$$CAM_{raw} = \sum_{k=0}^{K-1} w_k^c B(I)$$

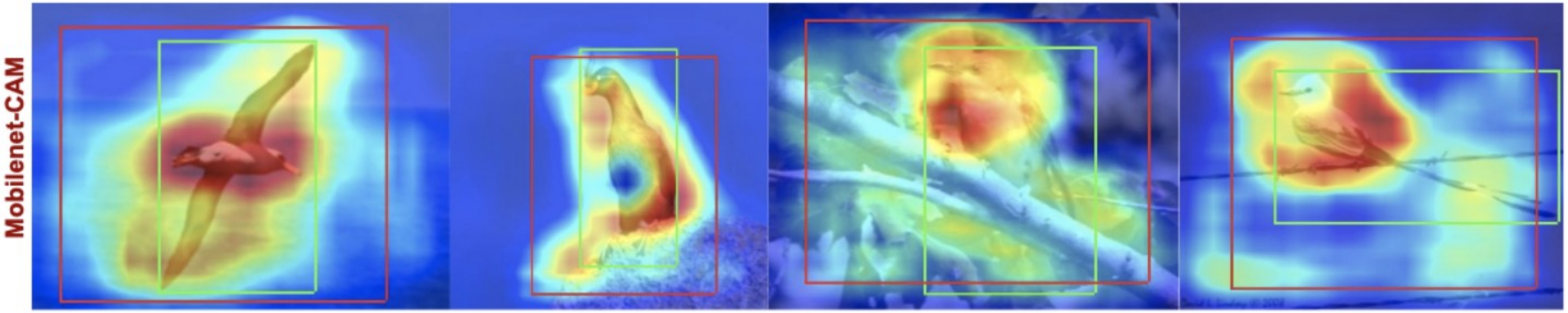
Weakly-supervised Localization via CAM

Annotation (GT)
Estimation

Some Problems



Under-estimation



Over-estimation

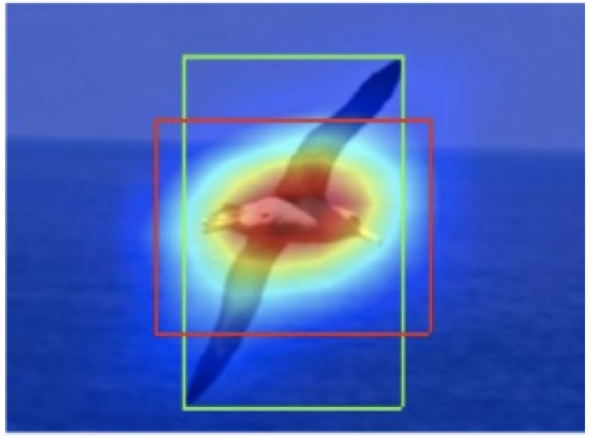
Related Work

Under-estimating Object Region

– Drop the most discriminative regions

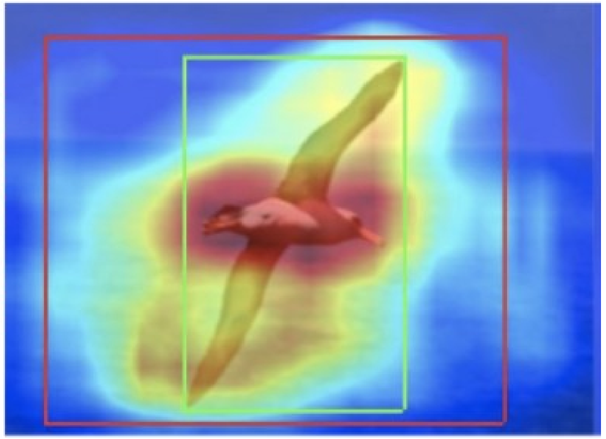
– Occlude parts of the input

[1] Singh et al., ICCV 17
[2] Choe et al., CVPR 19
[3] Zhang et al., CVPR 18
[4] Yang et al., WACV 20



Over-estimating Object Region

– Compute all possible CAMs and combine them via a pre-defined function.

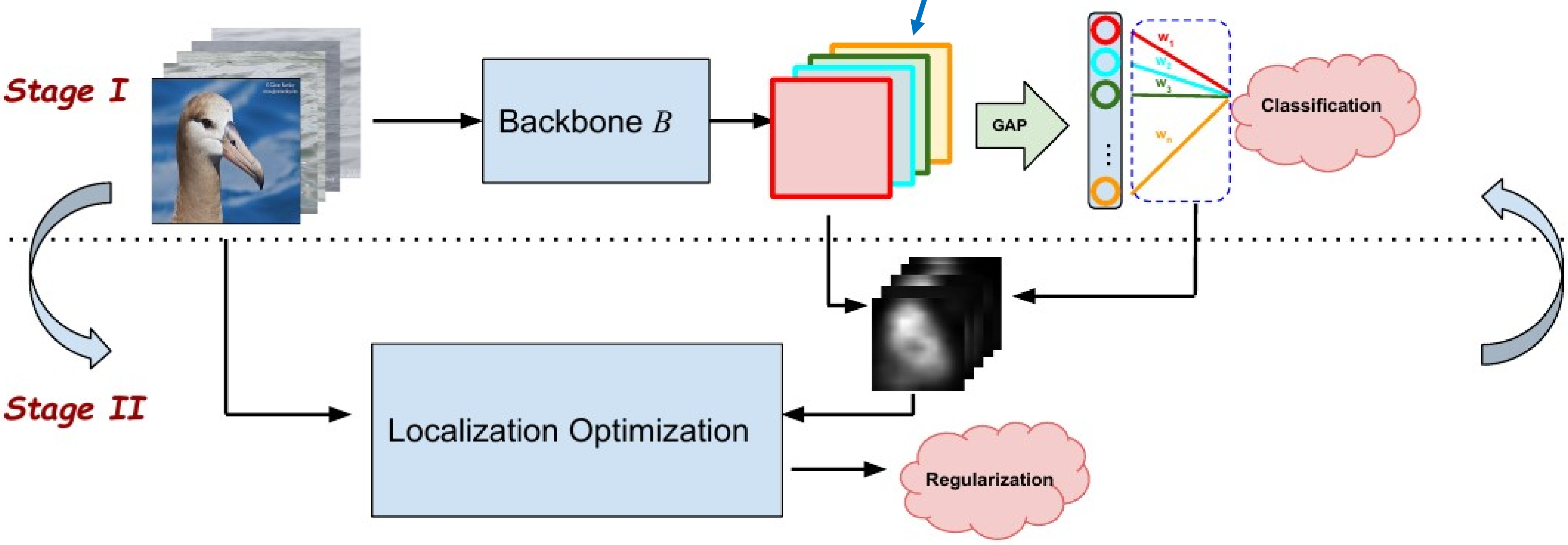


University of Antwerp imec

Proposed Method

$$CAM = \sum_{k=0}^{k-1} w_k^c B(I)$$

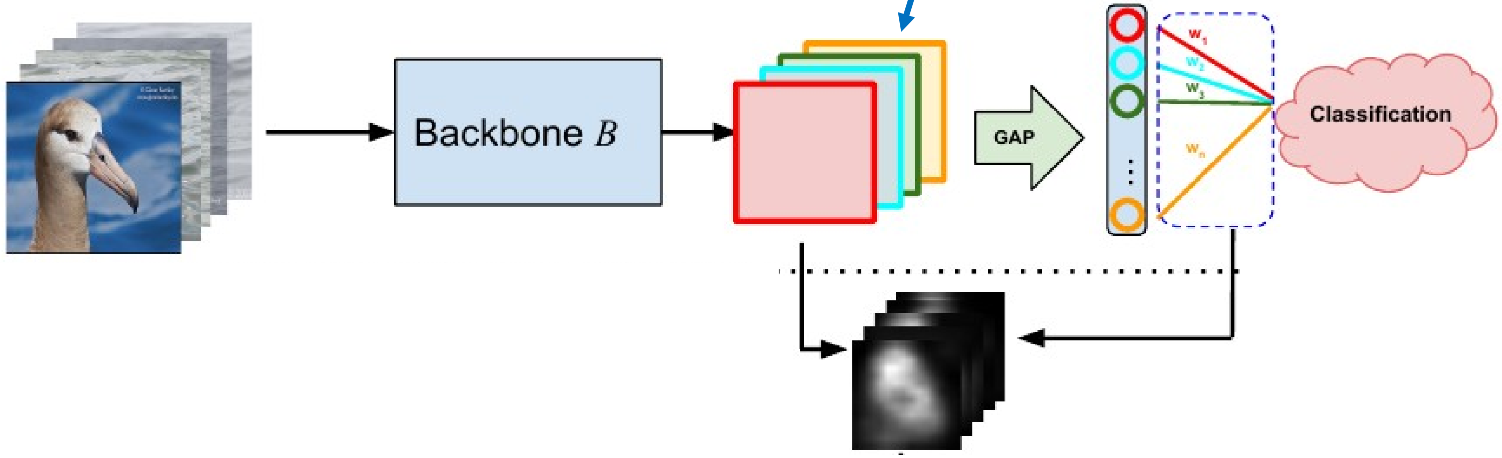
MinMaxCAM



Proposed Method

$$CAM = \sum_{k=0}^{k-1} w_k^c B(I)$$

MinMaxCAM – test time



MinMaxCAM [Wang et al., BMVC 2021]

Over-estimation

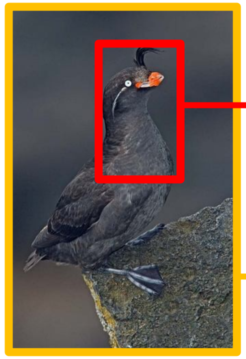
Common Region Regularization (CRR)

$$f = B(I * H)$$

$$CRR = \frac{1}{S(S-1)} \sum_{i=0}^{S-1} \sum_{j=0}^{S-1} ||f_i - f_j||_2^2$$

Under-estimation

Full Region Regularization (FRR)



$$f = B(I * H)$$

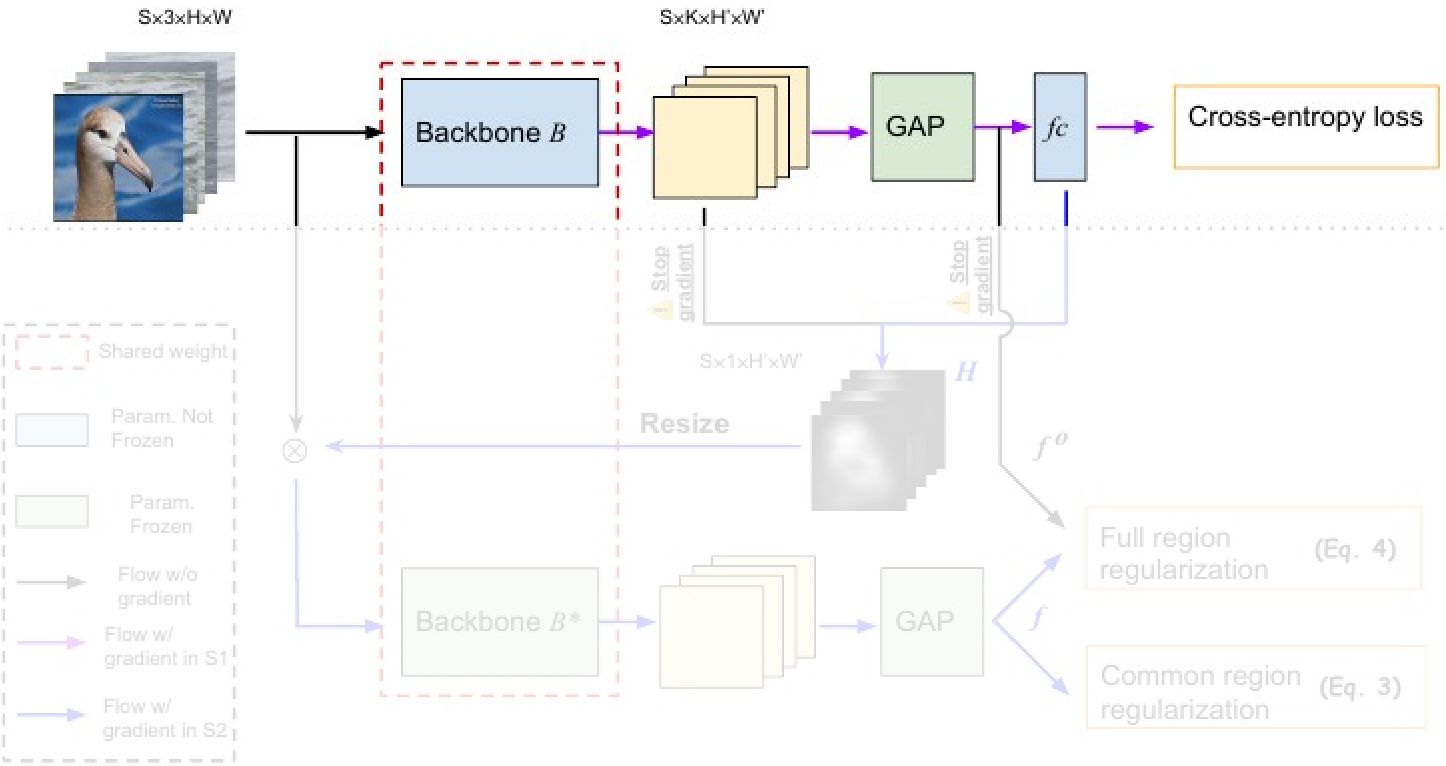
$$f^o = B(I)$$

$$FRR = \frac{1}{S} \sum_{i=0}^{S-1} ||f_i - f_i^o||_2^2$$

MinMaxCAM [Wang et al., BMVC 2021]

Training

→ Apply both loss functions per minibatch in an iterative manner



Stage-I

$$L_{S1} = -\sum_{i=1}^{N \times S} c_i log(\hat{y}_i)$$

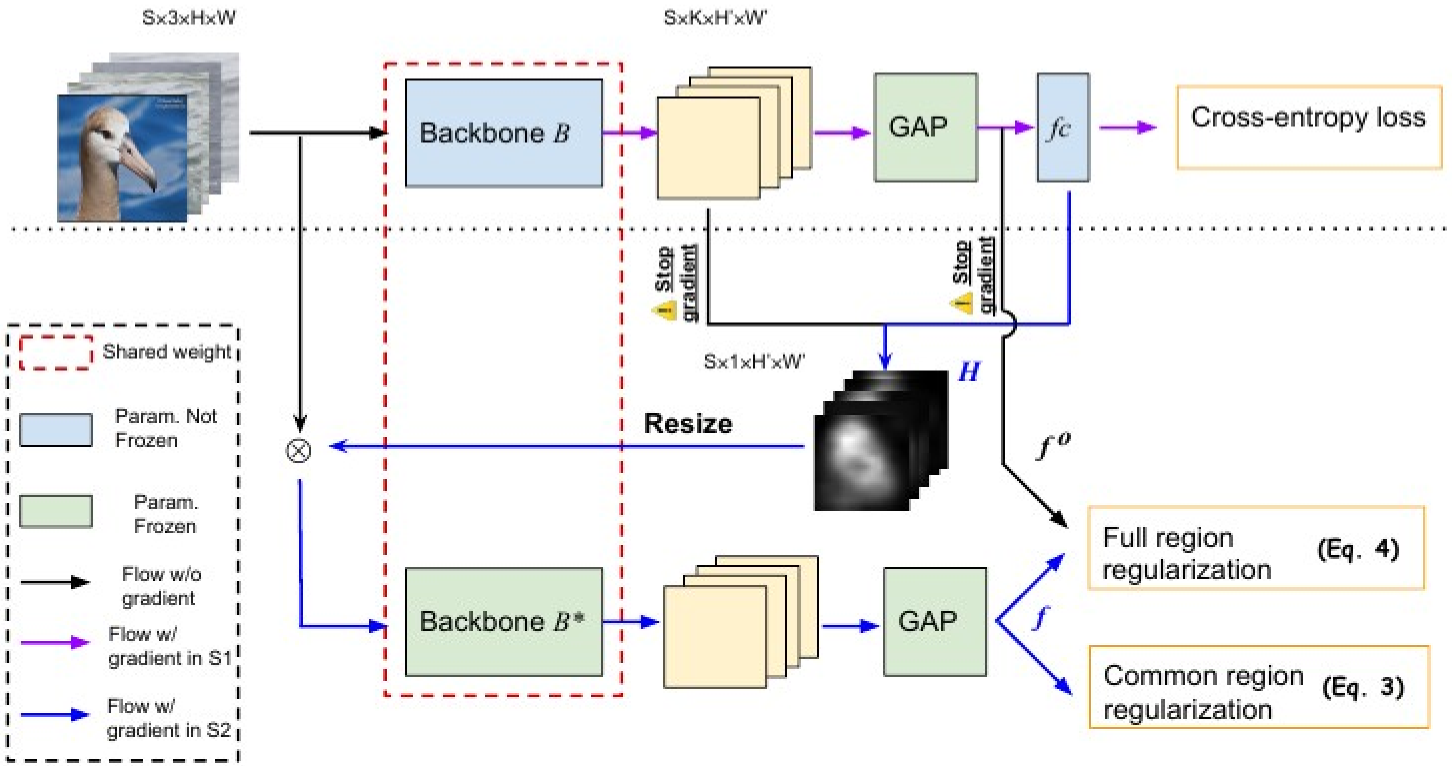
Stage-II

$$L_{S2} = \lambda_1 CRR + \lambda_2 FRR$$

MinMaxCAM [Wang et al., BMVC 2021]

Training

→ Apply both loss functions per minibatch in an iterative manner



Stage-I

$$L_{S1} = -\sum_{i=1}^{N\times S} c_i log(\hat{y}_i)$$

Stage-II

$$L_{S2} = \lambda_1 CRR + \lambda_2 FRR$$

Evaluation

Datasets

- ILSVRC'12 | CUB-200-Birds | OpenImages Segmentation

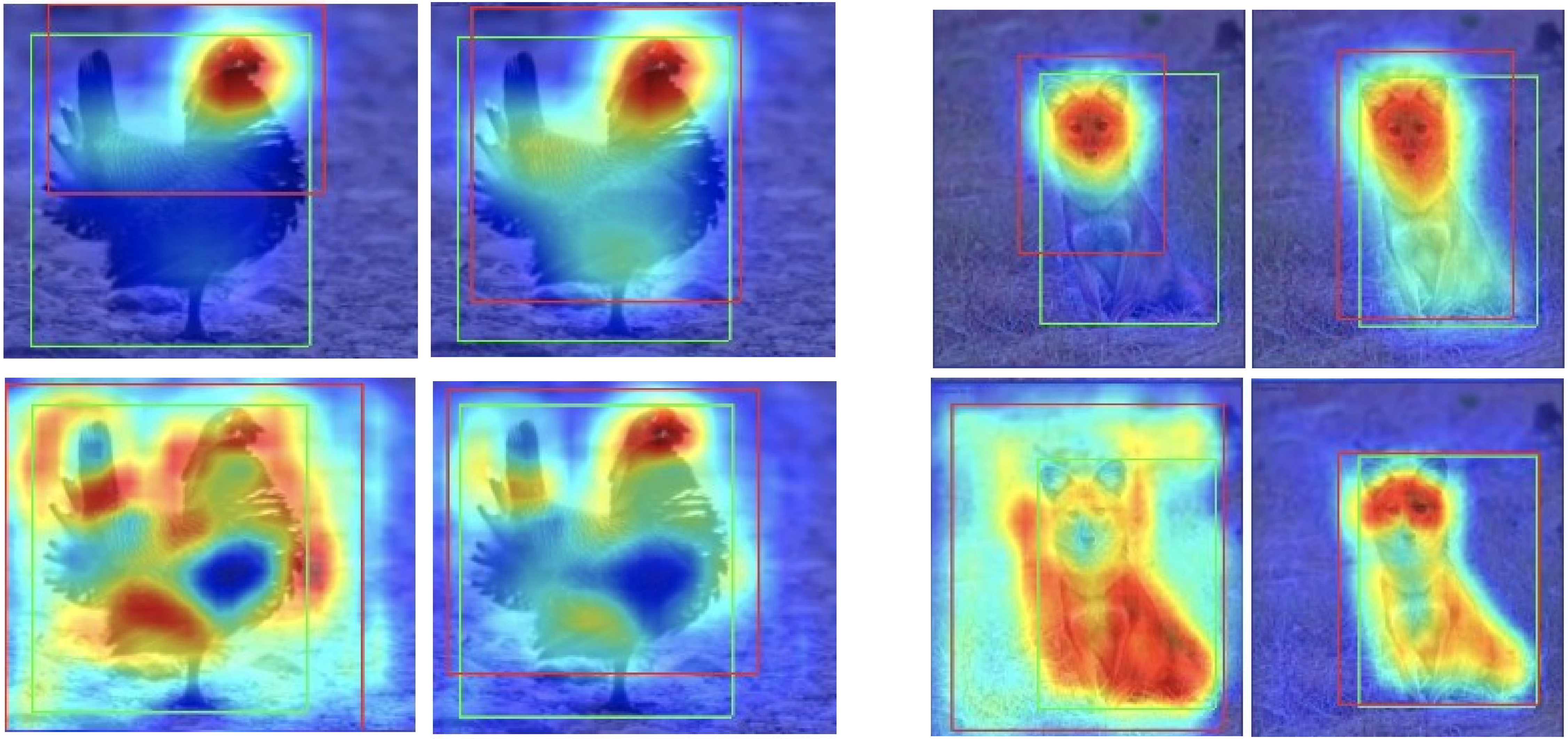
Architectures

- VGG-16 | ResNet-50 | MobileNetV2

Evaluation – Qualitative Results

Annotation (GT)
Estimation

Before After Before After

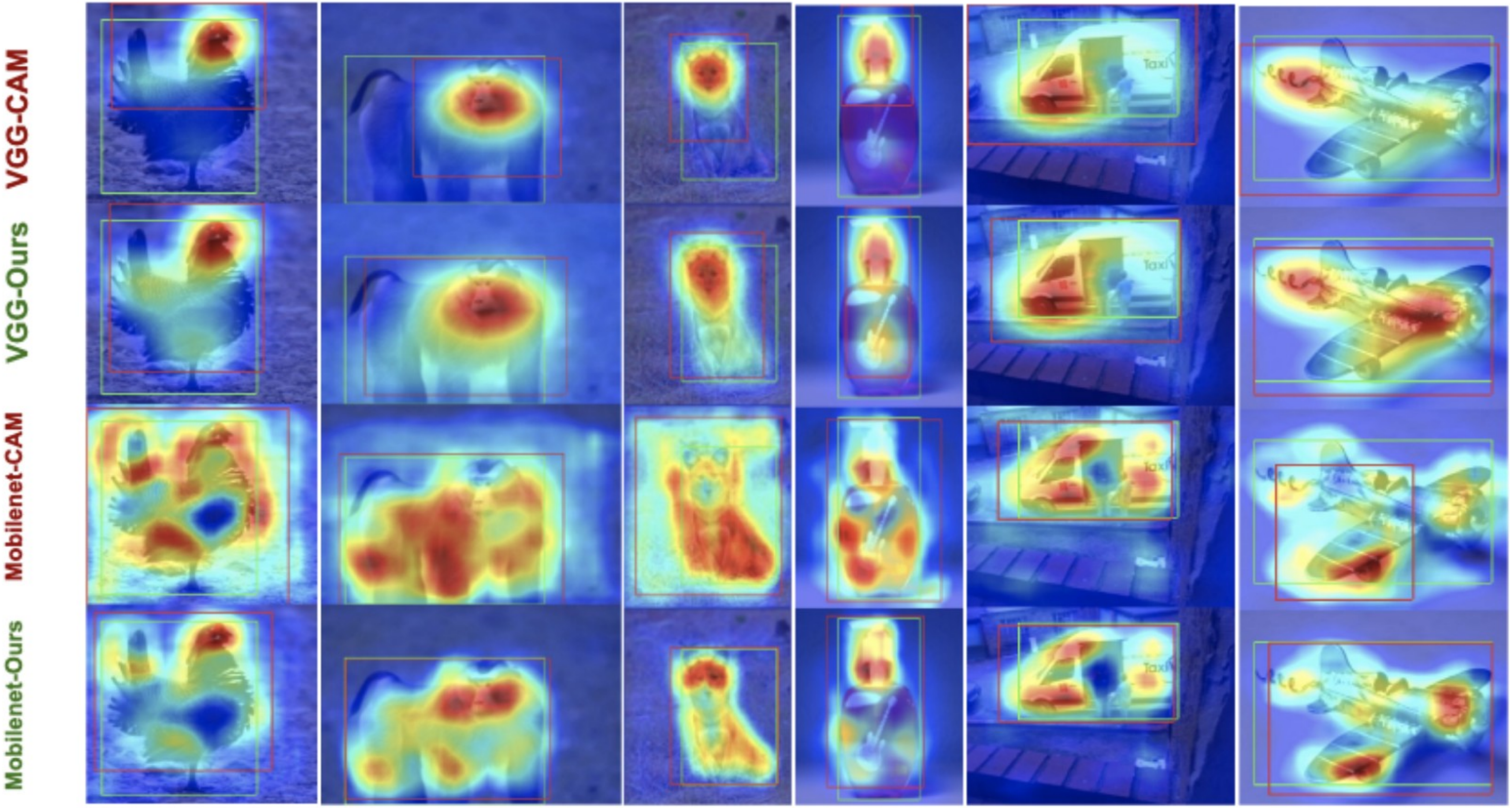


Under-estimation

Over-estimation

Evaluation – Qualitative Results

Annotation (GT)
Estimation



University of Antwerp

Evaluation – Quantitative Results

	Method	Backbone	ImageNet		CUB		OpenImages
			MaxBoxAcc (%)	MaxBoxAccV2 (%)	MaxBoxAcc (%)	MaxBoxAccV2 (%)	PxAP (%)
*	CAM	VGG16	61.1	60.0	71.1	63.7	58.1
	HaS	VGG16	0.7	0.6	5.2	0	-1.2
	ACoL	VGG16	-0.8	-2.6	1.2	-6.3	-3.4
	SPG	VGG16	0.5	-0.1	-7.4	-7.4	-2.2
	ADL	VGG16	-0.3	-0.2	4.6	2.6	0.2
	CutMix	VGG16	1.0	-0.6	0.8	-1.4	0.1
	I2C	VGG16	-	-	-2.7	-3	-1
	Ours	VGG16	***3.5***	***2.2***	***12.8***	***6.5***	***1.9***
*	CAM	ResNet50	64.2	63.7	73.2	63.0	58.0
	HaS	ResNet50	-1	-0.3	**4.9**	1.7	0.2
	ACoL	ResNet50	-2.5	-1.4	-0.5	3.5	-0.2
	SPG	ResNet50	-0.7	-0.4	-1.8	-2.6	-0.3
	ADL	ResNet50	0	0	0.3	-4.6	-3.7
	CutMix	ResNet50	-0.3	-0.4	-5.4	-0.2	-0.7
	I2C	ResNet50	-	-	0.3	1.0	**2.9**
	Ours	ResNet50	***2.5***	***2.0***	*4.8*	***4.3***	***2.9***
*	CAM	MobilenetV2	60.8	59.5	65.3	58.1	54.9
	I2C	MobilenetV2	-	-	1.9	1.5	3.3
	Ours	MobilenetV2	***4.5***	***3.8***	***10.5***	***6.9***	***4.4***

Evaluation – Quantitative Results

Ablation Study – Effect of the set size *S*

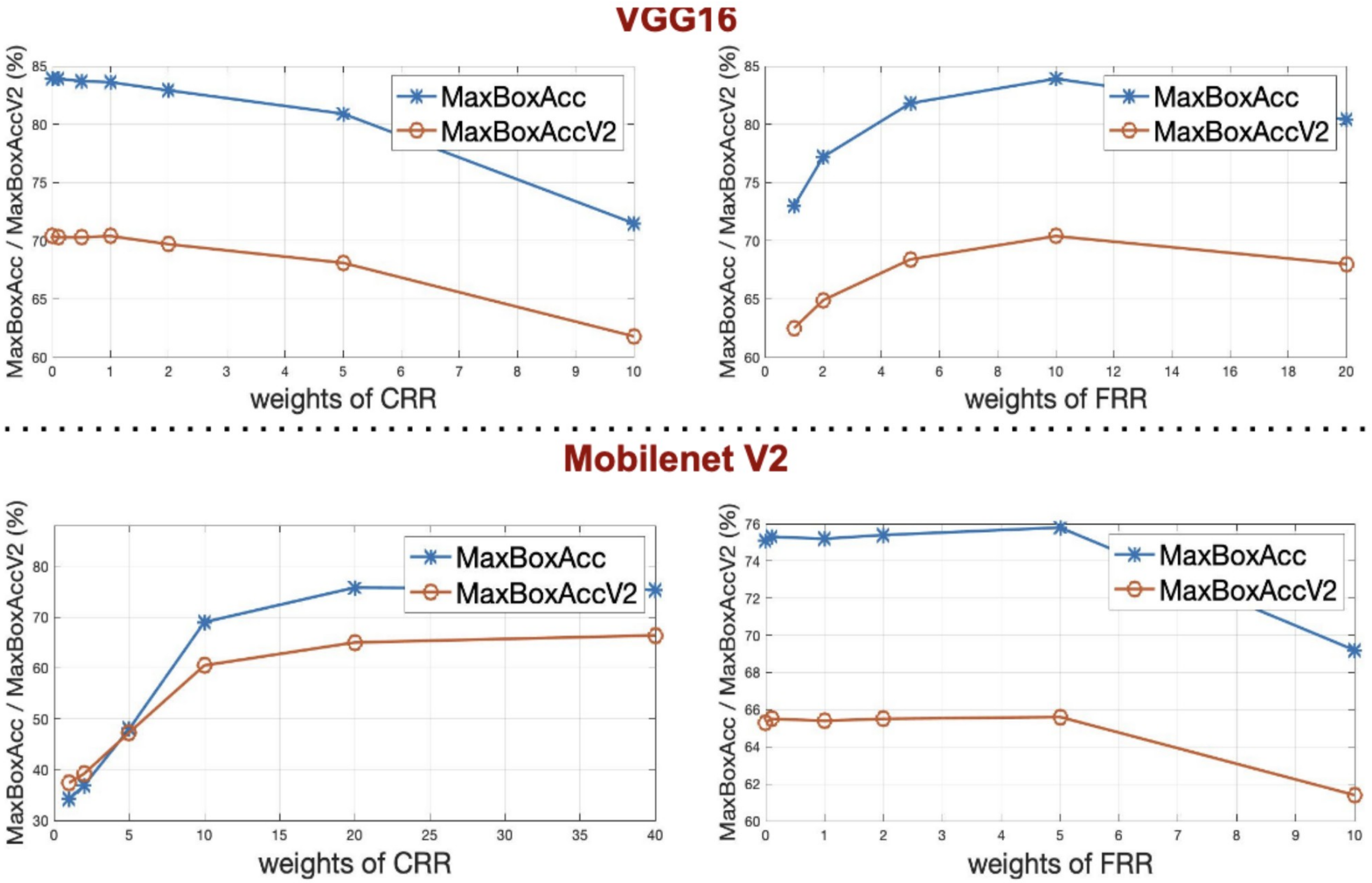
Set size S	$MaxBoxAcc(\%)$	$MaxBoxAccV2(\%)$
S=5	75.8	65.0
S=4	74.7	64.6
S=3	74.4	64.4
S=2	73.9	63.9
CAM	65.3	58.1

Table 3: Effect of the set size S.

Evaluation – Quantitative Results

Ablation Study – Effect of the hyperparameters λ

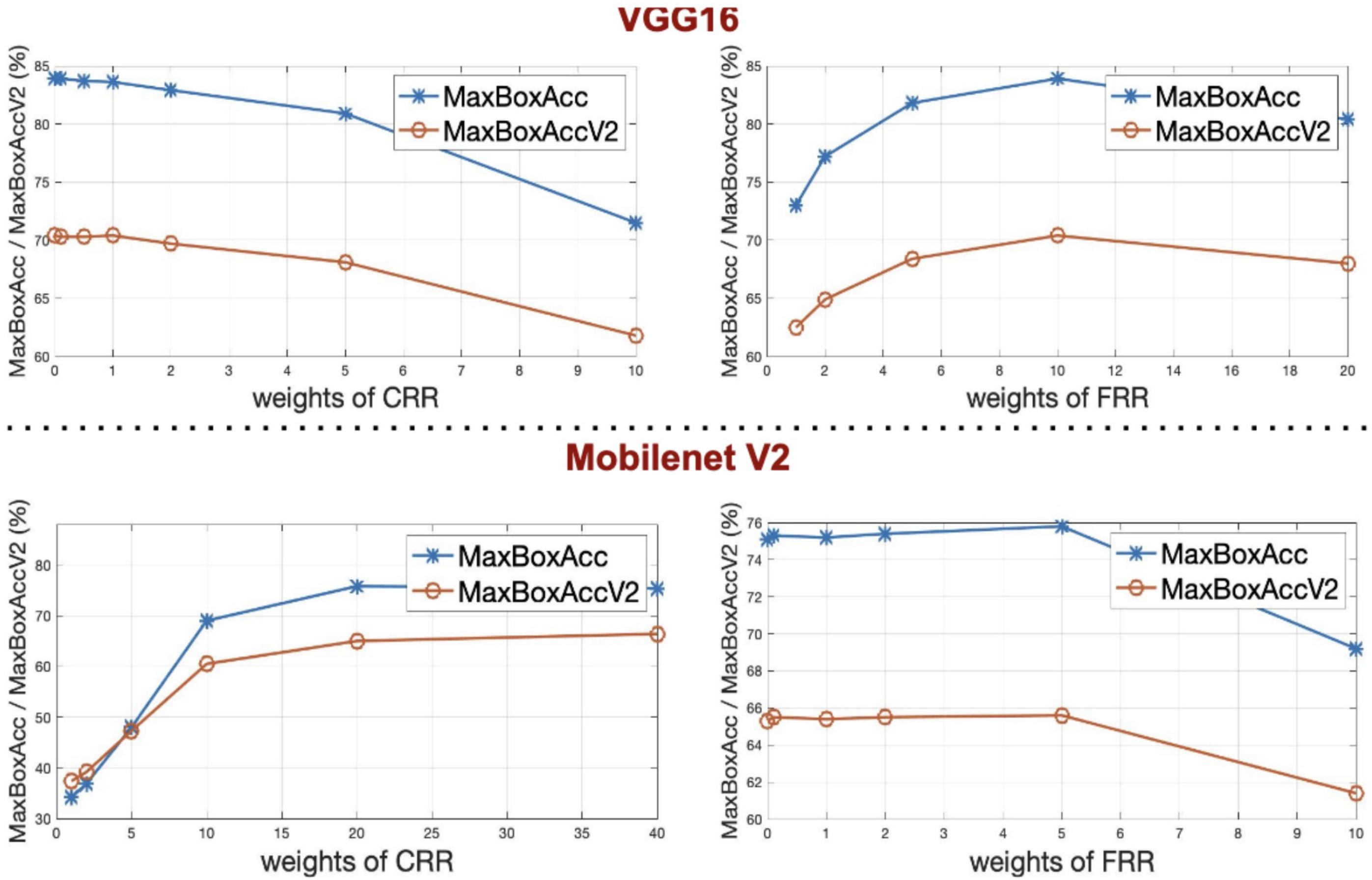
$$L_{S2} = \lambda_1 CRR + \lambda_2 FRR$$



Evaluation – Quantitative Results

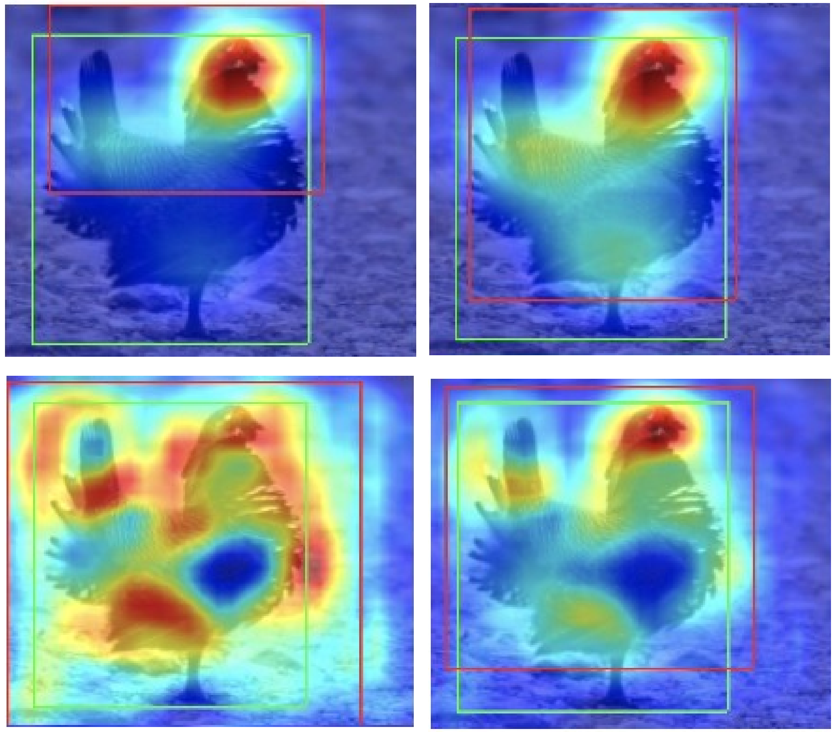
Ablation Study – Effect of the hyperparameters λ

$$L_{S2} = \lambda_1 CRR + \lambda_2 FRR$$

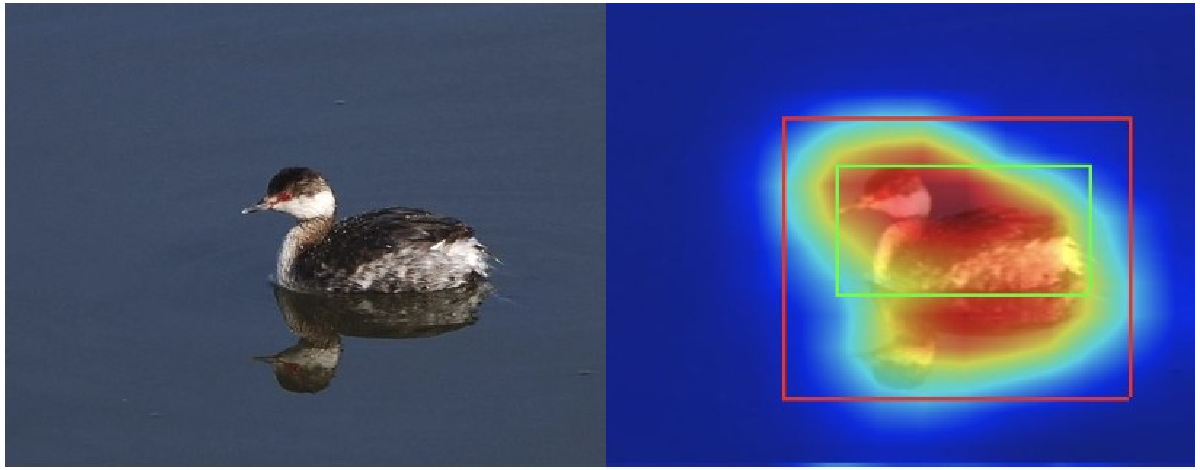


Before

After

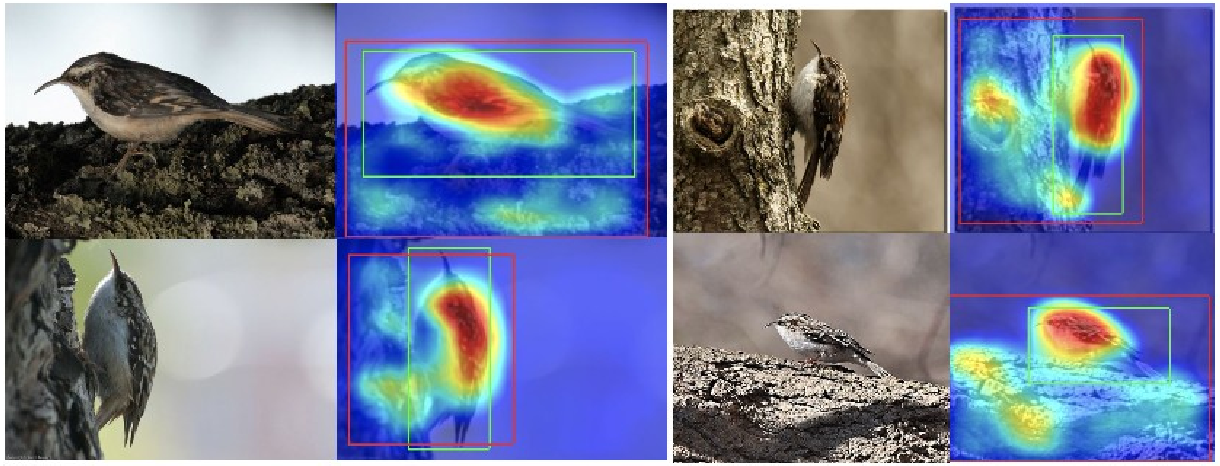


Evaluation – Failure Cases



Type 1 failure

Annotation (GT)
Estimation



Type 2 failure

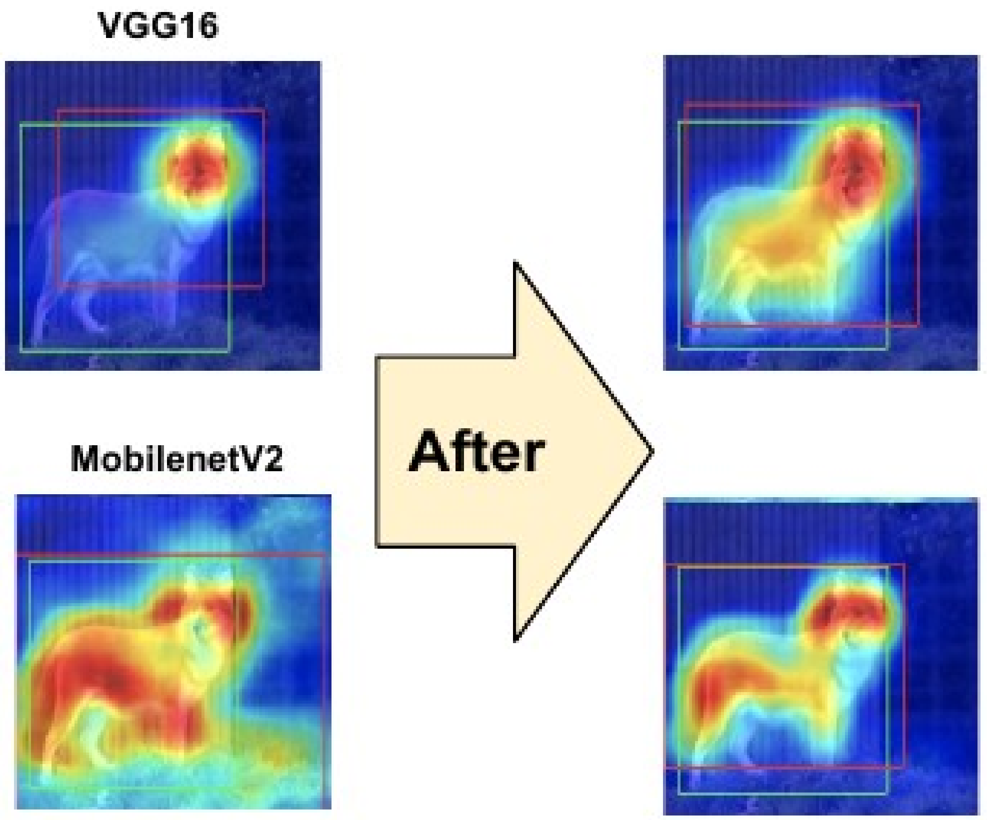
Take Home Message

Take Home Message

- MinMaxCAM
  - Redistribute the activation mass
  - Lightweight, Fast
  - Relatively simple to train
  - Limited to sigle-instance occurrence
- More details or results?
  - Please see the paper:

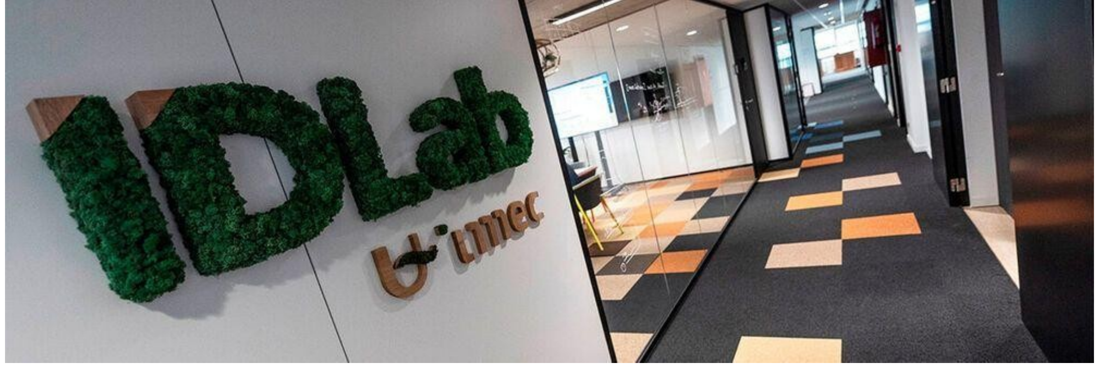
    MinMaxCAM: Improving object coverage for CAM-based Weakly Supervised Object Localization

    Kaili Wang, José Oramas M., and Tinne Tuytelaars. BMVC 2021
  - Publicly available [arxiv:2104.14375]



Thanks for your Attention

Want to discuss further?



José Oramas

Assistant Professor

Internet Data Lab, University of Antwerp & imec

Research Interests

- Representation Learning & Computer Vision
- Explainable AI / Interpretable ML

Contact details:

- **Email:** Jose Oramas@UAntwerpen.be
- **Twitter:** @jaom7
- **Mastodon:** sigmoid.social/@jaom7

Weakly-supervised object localization via class activation mapping

José Oramas

Internet Data Lab (IDLab), University of Antwerp, imec.

University of Antwerp imec