University of Trento [1]
Fondazione Bruno Kessler [2]
Mapillary Research [3]

Increasingly specialized ensemble of Convolutional Neural Networks for Fine-grained Recognition

Andrea Simonelli[1], Stefano Messelodi[2], Francesco De Natale[1], Samuel Rota Bulo’[3]

Overview

1. **Introduction**
    - Fine-grained recognition

2. **State of the art**
    - Common issues
    - Common solutions

3. **Proposed method**
    - Increasingly specialized ensemble of Convolutional Neural Networks for Fine-grained recognition

4. **Conclusions and future work**

1. Introduction

Fine-grained recognition

Discriminate among classes with **subtle differences**

“Standard”
classification task

High
inter-class variations

vs

Fine-grained
classification task

Small
inter-class variations

Why is this important?

To build systems able to solve **increasingly complex** tasks

2. State of the art

Common issues

- **Dataset size:** fine-grained datasets are usually small

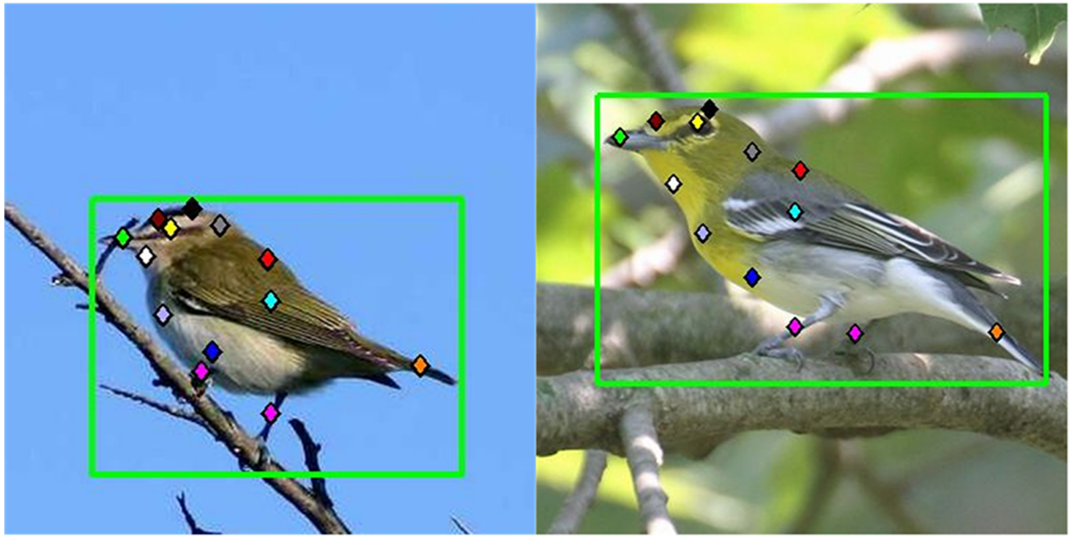
- **Inter-class variations:** on top of being **subtle** they can be **very localized**


Due to these major issues networks suffer of **severe overfitting**

Common solutions (1)

State-of-the-art methods usually **combine localization** with **classification**:

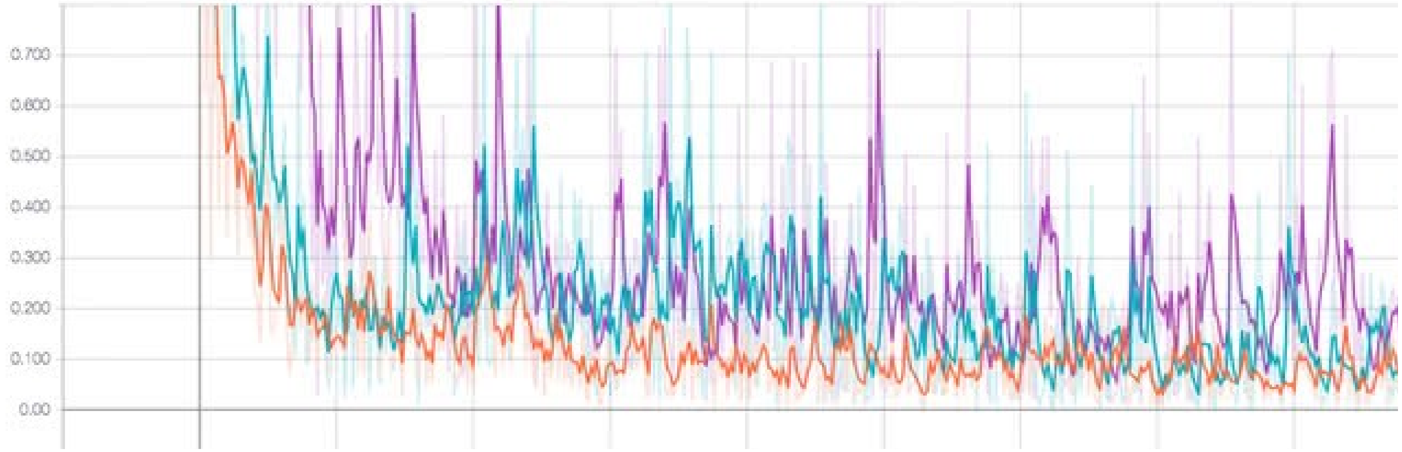
- **Fully supervised methods** rely on **annotations** like object or parts location



Annotations can be **very expensive** to obtain

Common solutions (2)

- **Weakly supervised methods** instead **learn** where discriminative parts are without annotations



Usually adopting **multiple losses**, **many extra hyper-parameters** requiring a **complex training procedure**

3. Proposed solution

Opening the black box of CNNs

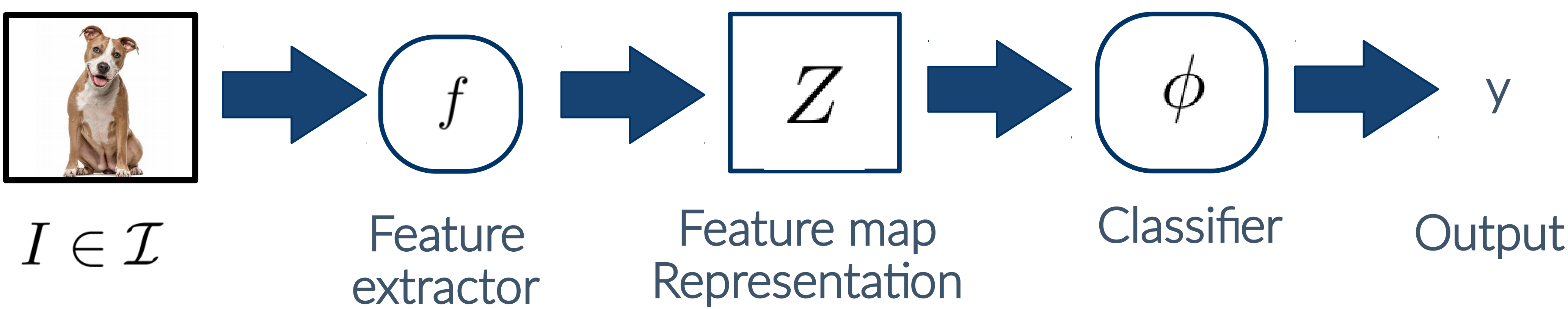
A **CNN** can be seen as a function **g(•)** which is the **composition** of:

- **Feature extractor f(•) detects features** and creates a ***representation Z***

$$f : \mathcal{I} \rightarrow \mathcal{Z}$$

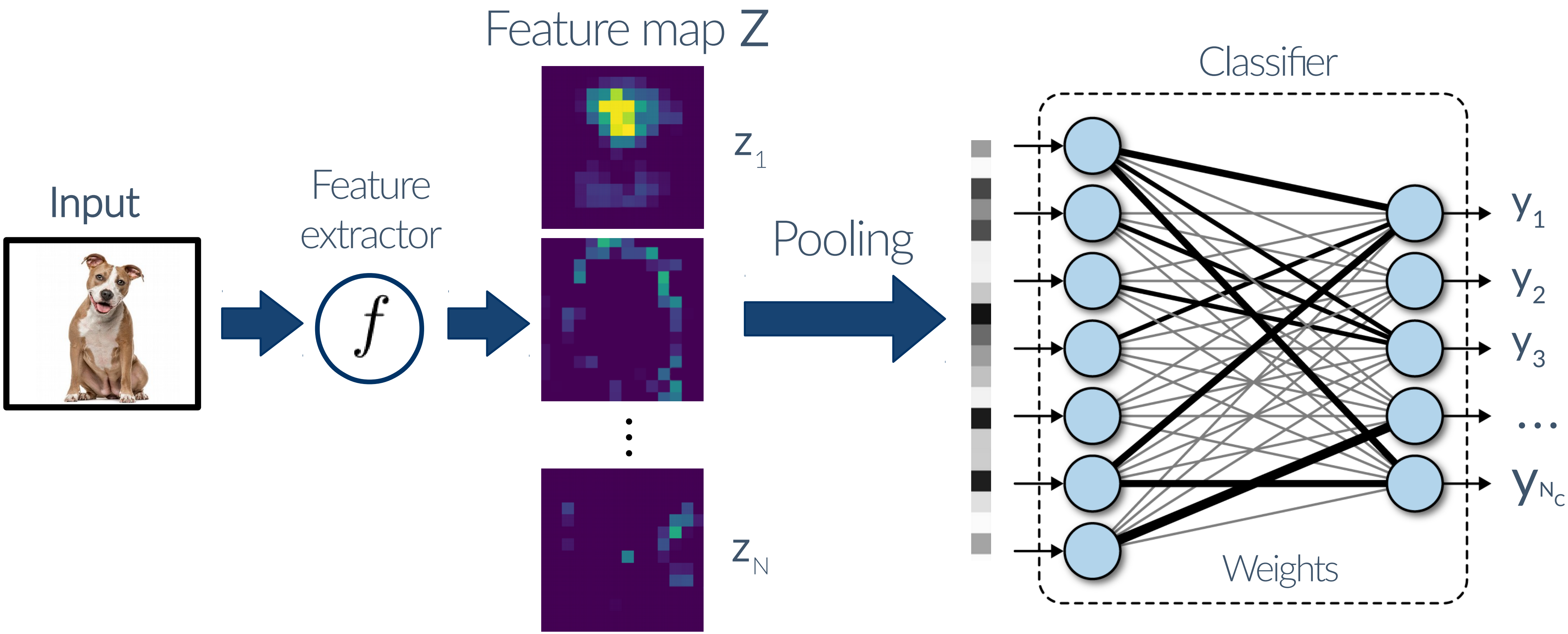
- **Classifier ϕ(•)** which **combines features** in **Z** to **predict** output **y**

$$\phi : \mathcal{Z} \rightarrow \mathcal{Y}$$



Looking closer

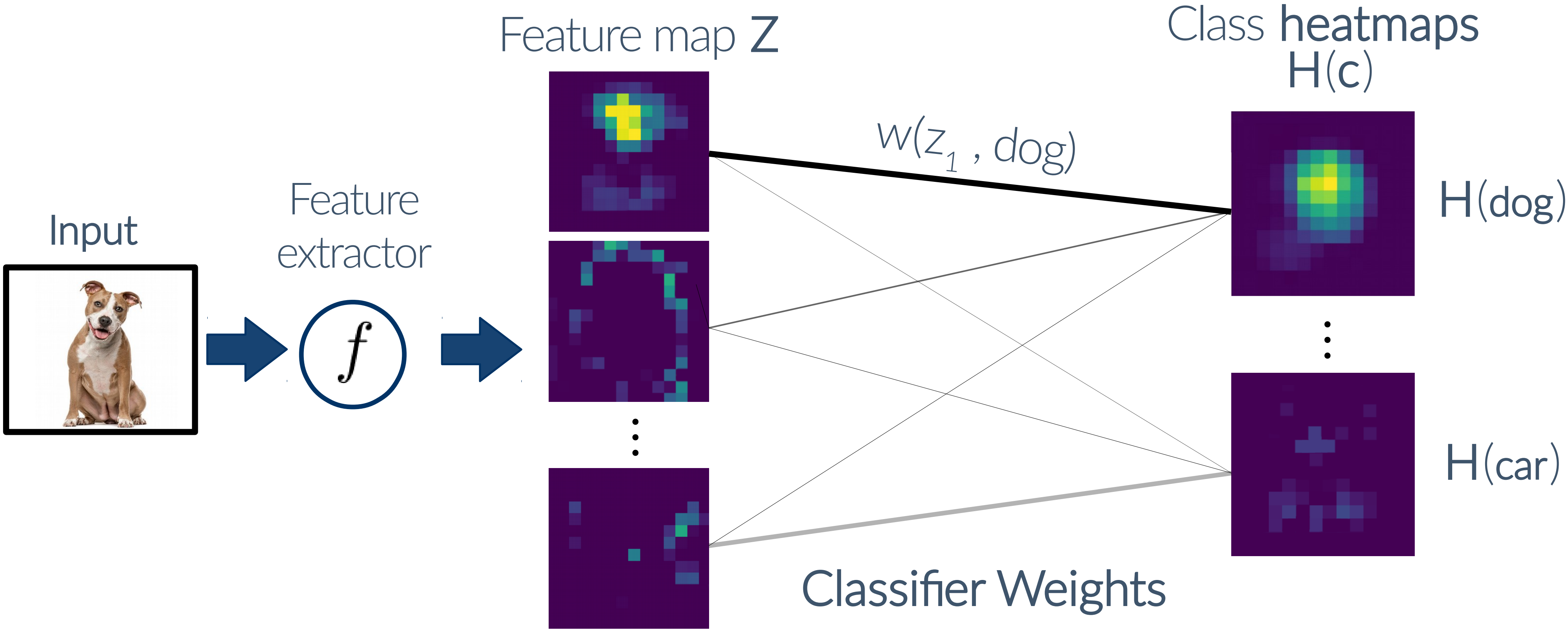
- Feature maps encode the **presence of features** in **specific regions**
- Classifier **combines** and **weights** (pooled) features to compute the outputs



Classifier weights encode the **importance** of **each feature** for **each class**

A nearly-free localization

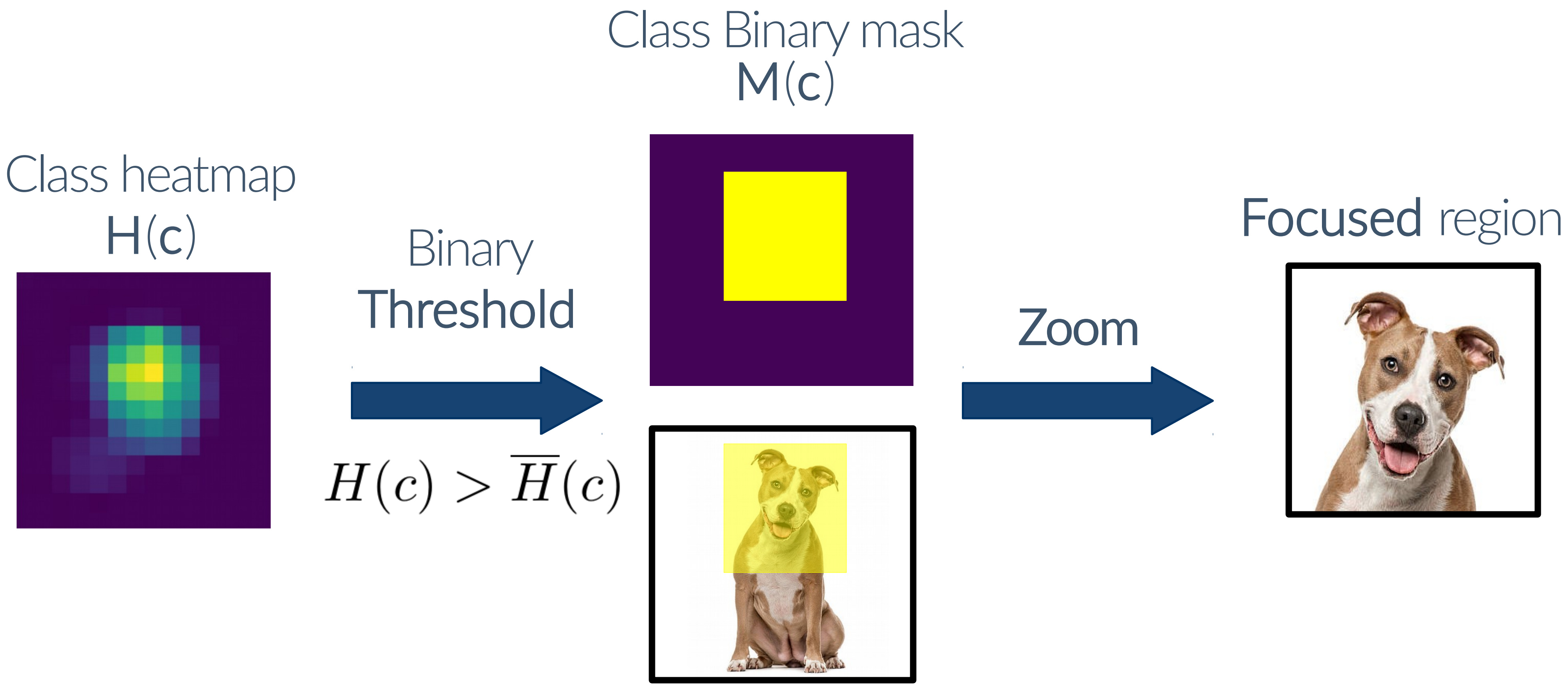
Zhang et al.[1] proposed to weight features **preserving the spatial information**



This results in **class heatmaps** where "high" pixels contain **class features**

Focus operation

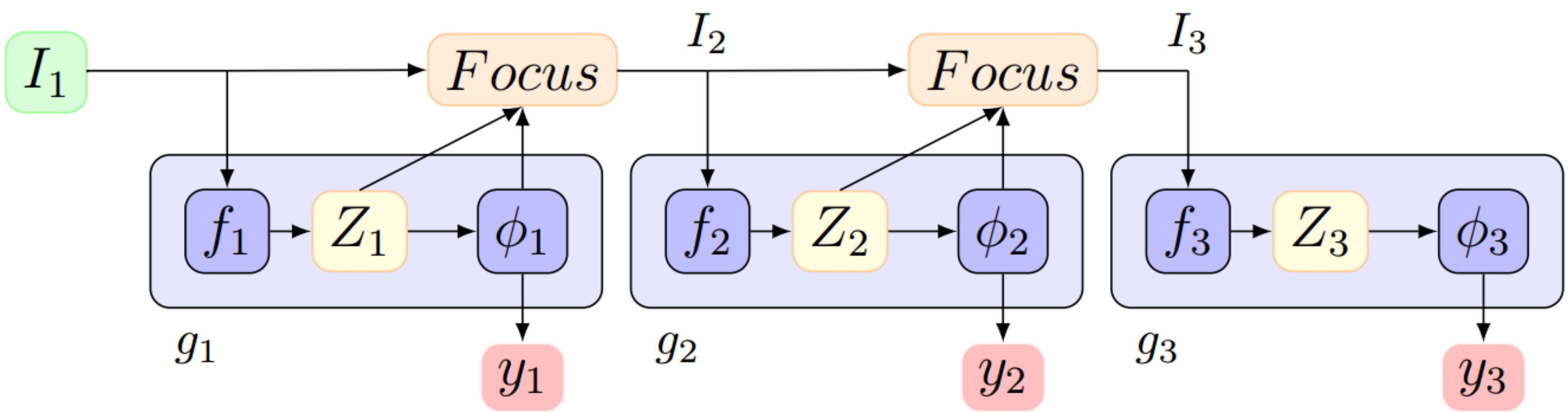
Applies a **binary threshold** to class heatmaps to select a **relevant region**



The region above threshold is **extracted** and **zoomed** to find **finer details**

Combine CNNs in an ensemble

This **focus** operation is performed **between** consecutive **CNNs**



Coarse input

First CNN

Focus

"Finer" input

Second CNN

Focus

"Finest" input

Last CNN

The networks achieve **increasingly higher** level of **specialization**

Results

Let's now compare the ensemble with current **state-of-the art** methods:

CUB-Birds [2]

200 species of **birds**
~6k training images

Method	Annotations	Accuracy
Part-RCNN	✓	76.4
FCAN	✓	84.7
Zhang et al.		84.7
RA-CNN		85.3
Resnet-50		85.5
DT-RAM		86.0
MA-CNN		86.5
Ours		**87.2**

FGVC-Aircraft [3]

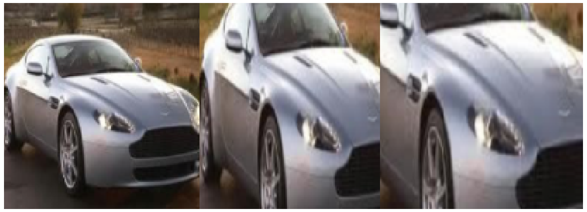
100 types of **airplanes**
~6k training images

Method	Annotations	Accuracy
Zhang et al.		87.3
RA-CNN		88.2
Resnet-50		89.0
MA-CNN		89.9
Ours		**90.9**

Stanford Cars [4]

196 **car** models
~10k training images

Method	Annotations	Accuracy
Zhang et al.		91.7
RA-CNN		92.5
MA-CNN		92.8
FCAN	✓	93.1
DT-RAM		93.1
Resnet-50		93.3
Ours		**94.1**



Ablation studies

Let's compare the accuracy of the **single networks** with the **ensemble**:

$Dataset$	y_1	y_2	$\hat{y}_2$	y_3	$\hat{y}_3$
CUB-200 [1]	85.5	83.4	86.8	83.6	87.2
FGVC-Air. [2]	89.0	88.6	90.6	87.3	90.9
Stanf. Cars [3]	93.3	92.7	94.0	91.1	94.1

Where y_n is the performance of the **single** network at the **n**th stage of the ensemble and $\hat{y}_N$ is the accuracy of the **ensemble** with **N** networks

The accuracy of the ensemble **always exceeds** the one of the single network

4. Conclusions and future work

Conclusions and future work

The proposed method:

- Is **simple**
- Achieves **state-of-the-art results** on three popular fine-grained datasets
- Does **not** require extra hyper-parameter tuning, training or annotations

Future work will be geared towards the definition of a **recurrent model** as well as to the application of this study in **real-world problems**

Implementation

- **Architecture:** Resnet-50[5] pre-trained on Imagenet[6]

- **Optimization:** SGD with momentum 0.9 for 270 epochs

- **Losses:** Cross Entropy loss

- **Learning rate:** initially 1e-3 later decreased by 1/10 every 100 epochs

- **Regularization:** dropout rate 0.7, L2 with decay 5e-4

- **Input sizes:** coarse input at 448x448px, others at 224x224px

- **Augmentations:** random {flips, resizing, crops, distortion (bright., contr., satur.)}

- **Framework:** implemented in Pytorch

Thank you!

References

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929.

[2] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.

[3] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," Tech. Rep., 2013.

[4] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in Proc. of the International Conference on Computer Vision Workshops (ICCVW), 2013, pp. 554–561.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1097–1105.