Return of the Devil in the Details:
Delving Deep into Convolutional Nets

Ken Chatfield - Karen Simonyan - Andrea Vedaldi - Andrew Zisserman
University of Oxford

The Devil is still in the Details

2011

The devil is in the details: an evaluation of recent feature encoding methods

Ken Chatfield
http://www.robots.ox.ac.uk/~ken
Victor Lempitsky
http://www.robots.ox.ac.uk/~vilem
Andrea Vedaldi
http://www.vlfeat.org/~vedaldi
Andrew Zisserman
http://www.robots.ox.ac.uk/~az

Department of Engineering Science,
Oxford University

Abstract

A large number of novel encodings for bag of visual words models have been proposed in the past two years to improve on the standard histogram of quantized local features. Examples include locality-constrained linear encoding [23], improved Fisher encoding [17], super vector encoding [27], and kernel codebook encoding [20]. While several authors have reported very good results on the challenging PASCAL VOC classification data by means of these new techniques, differences in the feature computation and learning algorithms, missing details in the description of the methods, and different tuning of the various components, make it impossible to compare directly these methods and hard to reproduce the results reported. This paper addresses these shortcomings by carrying out a rigorous evaluation of these new techniques by: (1) fixing the other elements of the pipeline (features, learning, tuning); (2) disclosing all the implementation details, and (3) identifying both those aspects of each method which are particularly important to achieve good performance, and those aspects which are less critical. This allows a consistent comparative analysis of these encoding methods. Several conclusions drawn from our analysis cannot be inferred from the original publications.

1 Introduction

The typical object recognition pipeline is composed of the following three steps: (i) extraction of local image features (*e.g.*, SIFT descriptors), (ii) encoding of the local features in an image descriptor (*e.g.*, a histogram of the quantized local features), and (iii) classification of the image descriptor (*e.g.*, by a support vector machine). Recently several authors have focused on improving the second component, *i.e.* the encoding of the local features in global

2014

Return of the Devil in the Details: Delving Deep into Convolutional Nets

Ken Chatfield
ken@robots.ox.ac.uk
Karen Simonyan
karen@robots.ox.ac.uk
Andrea Vedaldi
vedaldi@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford
Oxford, UK

Abstract

The latest generation of Convolutional Neural Networks (CNN) have achieved impressive results in challenging benchmarks on image recognition and object detection, significantly raising the interest of the community in these methods. Nevertheless, it is still unclear how different CNN methods compare with each other and with previous state-of-the-art shallow representations such as the Bag-of-Visual-Words and the Improved Fisher Vector. This paper conducts a rigorous evaluation of these new techniques, exploring different deep architectures and comparing them on a common ground, identifying and disclosing important implementation details. We identify several useful properties of CNN-based representations, including the fact that the dimensionality of the CNN output layer can be reduced significantly without having an adverse effect on performance. We also identify aspects of deep and shallow methods that can be successfully shared. In particular, we show that the data augmentation techniques commonly applied to CNN-based methods can also be applied to shallow methods, and result in an analogous performance boost. Source code and models to reproduce the experiments in the paper is made publicly available.

1 Introduction

Perhaps the single most important design choice in current state-of-the-art image classification and object recognition systems is the choice of visual features, or image representation. In fact, most of the quantitative improvements to image understanding obtained in the past dozen years can be ascribed to the introduction of improved representations, from the *Bag-of-Visual-Words* (BoVW) [6, 28] to the *(Improved) Fisher Vector* (IFV) [23]. A common characteristic of these methods is that they are largely *handcrafted*. They are also relatively simple, comprising dense sampling of local image patches, describing them by means of visual descriptors such as SIFT, encoding them into a high-dimensional representation, and

Comparing Apples to Apples:
State-of-the-art back in 2011

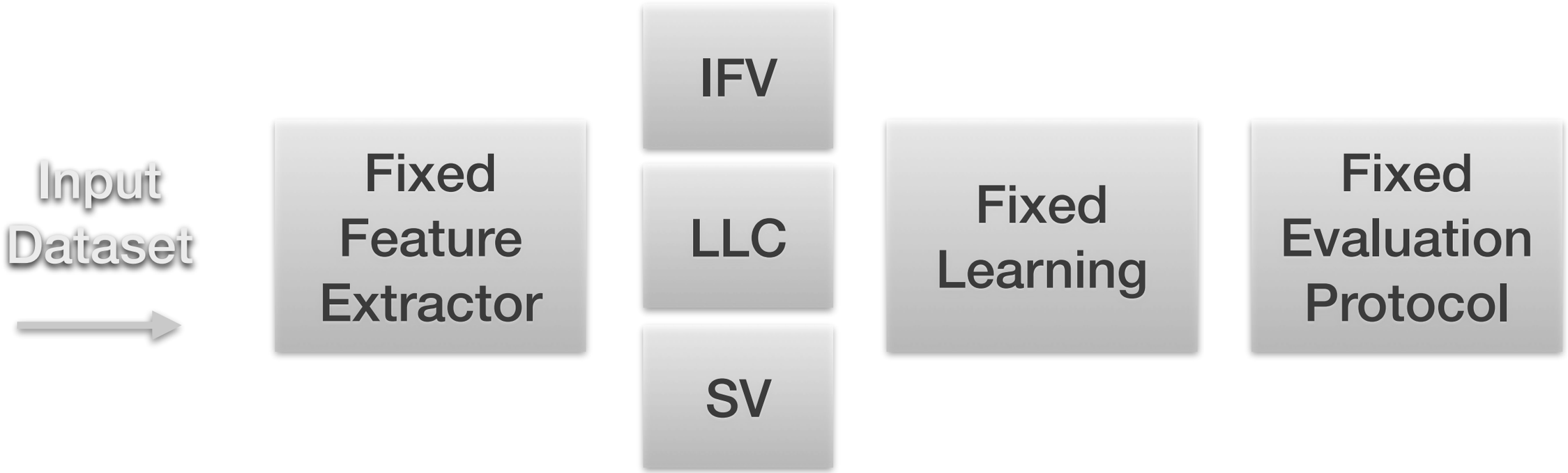
Back in 2011, state-of-the-art **image classification** pipelines were commonly based on the bag of visual words approach, with highly tuned feature encoders

IFV	LLC	SV
Improved Fisher Vector	Locality Constrained Linear Coding	Super-Vector Encoding

There were many **feature encodings** for this being proposed, but it was difficult to tell which worked best

Comparing Apples to Apples:
State-of-the-art back in 2011

In our previous work (BMVC 2011) we conducted an extensive **evaluation of these encodings** comparing them all on a **common-ground**:



* we'll call the features from these encodings **shallow** to distinguish them from the CNN-based features which follow

What's Changed?
State-of-the-art in 2014

- Introduction of CNN-based **deep visual features** to the community, all using pre-trained networks (Krizhevsky et al. 2012, Donahue et al. 2013, Oquab et al. 2014, Sermanet et al. 2014)

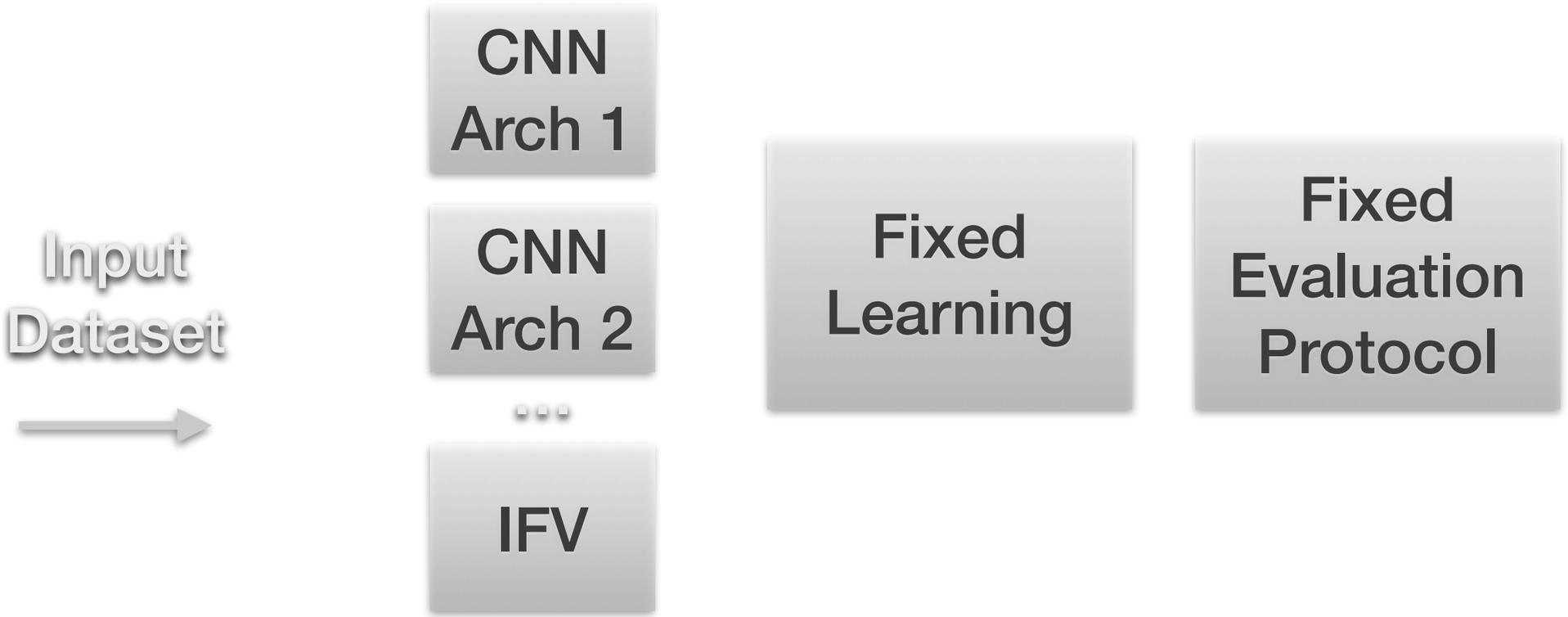
- Have shown to perform excellently over standard classification and detection benchmarks

- Unclear how the different methods introduced recently **compare to each other, and to shallow methods** such as IFV

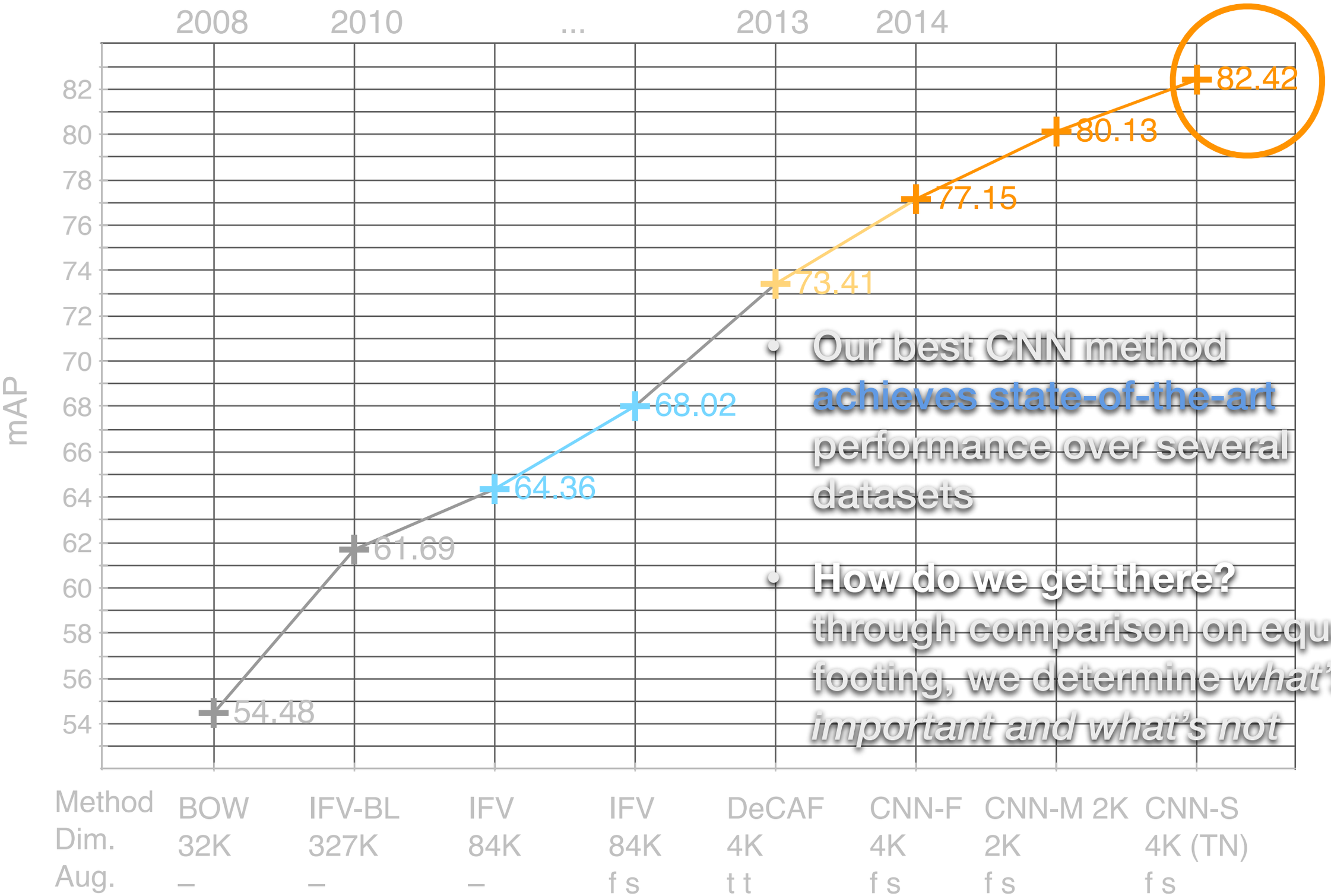
Comparing Apples to Apples:
State-of-the-art in 2014

- This work is again about **comparing the latest methods on a common ground**

- We compare both different **pre-trained network architectures** and different **learning heuristics**



Performance Evolution over VOC2007



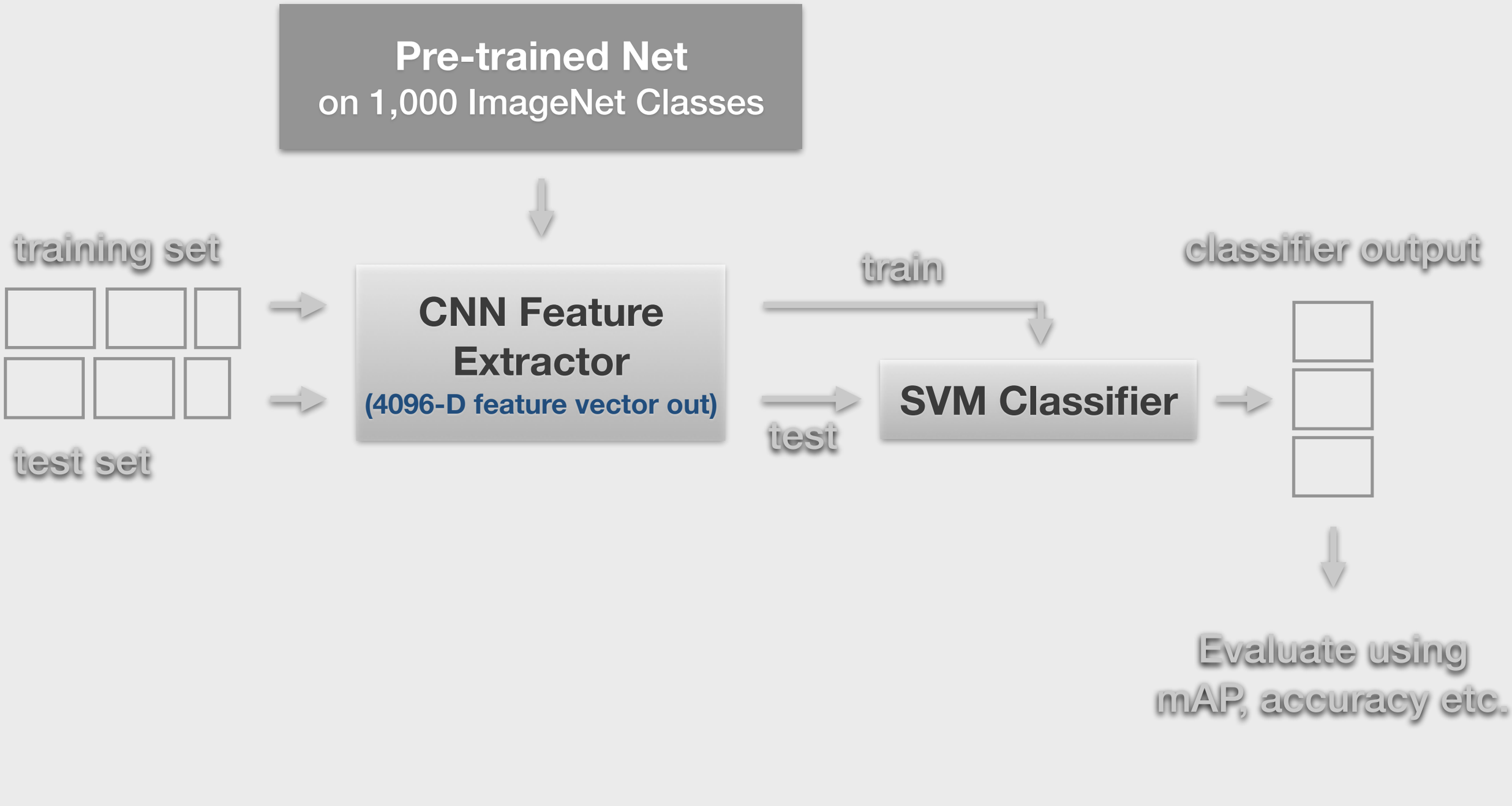
- **Our best CNN method achieves state-of-the-art performance over several datasets**
- **How do we get there? through comparison on equal footing, we determine *what's important and what's not***

Outline

Study Introduction and Evaluation Setup

1. Different pre-trained networks
2. Data augmentation (for both CNN and IFV)
3. Dataset fine-tuning
4. Reducing CNN final layer output dimensionality
5. Colour and CNN / IFV

Evaluation Setup



Pre-trained Networks

- **CNN-F** similar to Krizhevsky et al., NIPS 2012:
  'ImageNet classification with deep convolutional networks'

conv1 64x11x11 stride 4	conv2 256x5x5 stride 1	conv3 256x3x3 stride 1	conv4 256x3x3	conv5 256x3x3	fc6 4096 d.o.	fc7 4096 drop-out

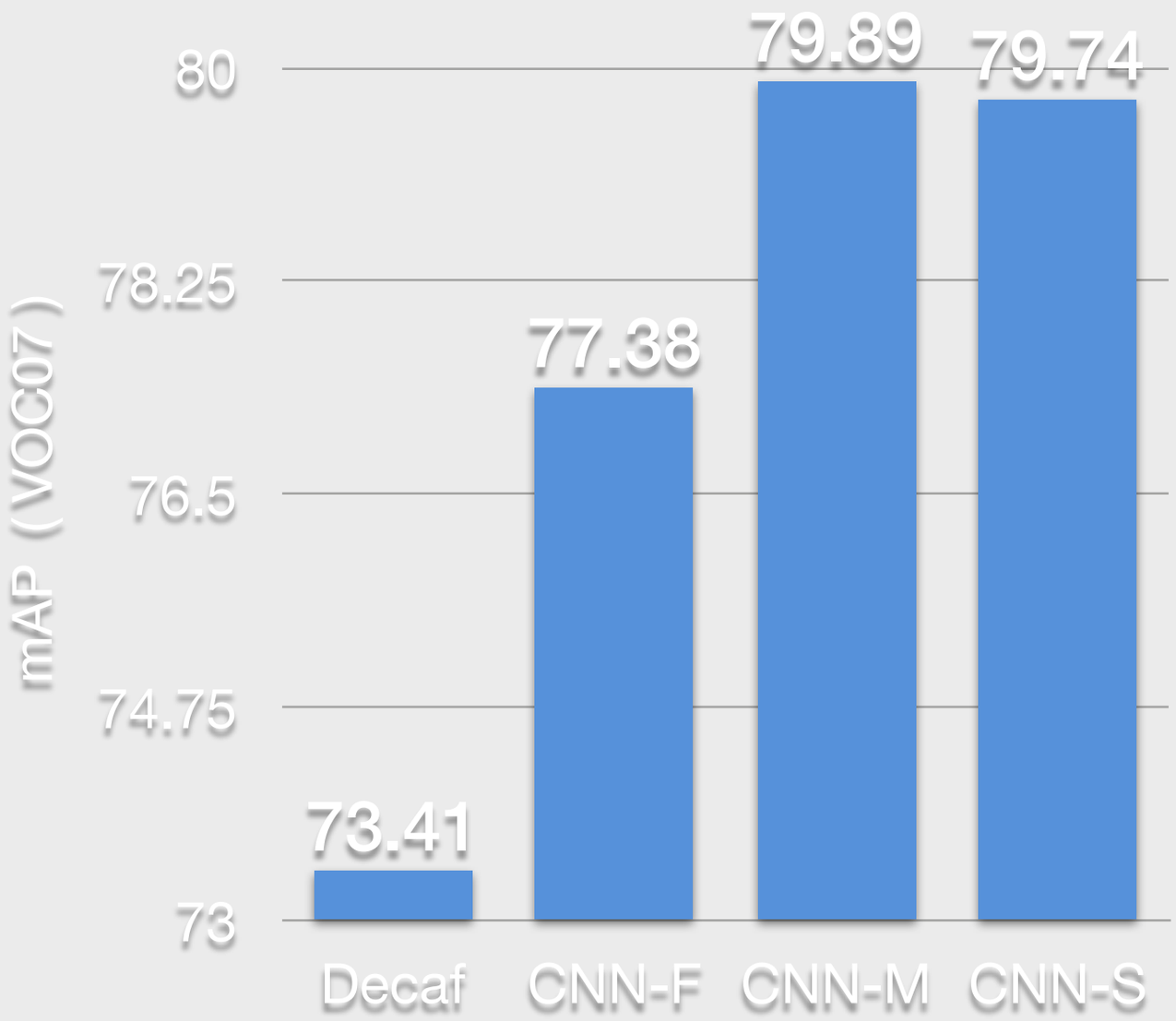
- **CNN-M** similar to Zeiler and Fergus, CoRR 2013:
  'Visualising and understanding convolutional networks'

conv1 96x7x7 stride 2	conv2 256x5x5 stride 2	conv3 512x3x3 stride 1	conv4 512x3x3	conv5 512x3x3	fc6 4096 d.o.	fc7 4096 drop-out

- **CNN-S** similar to OverFeat 'accurate' network, ICLR 2014:
  'OverFeat: integrated recognition, localisation and detection using ConvNets'

conv1 96x7x7 stride 2	conv2 256x5x5 stride 1	conv3 512x3x3 stride 1	conv4 512x3x3	conv5 512x3x3	fc6 4096 d.o.	fc7 4096 drop-out

Pre-trained Networks



Outline

Study Introduction and Evaluation Setup

1 Different pre-trained networks

2 Data augmentation (for both CNN and IFV)

3 Dataset fine-tuning

4 Reducing CNN final layer output dimensionality

5 Colour and CNN / IFV

Data Augmentation

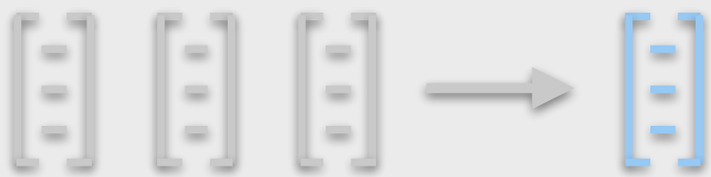
What do we mean by data augmentation?

Network Pre-training
(with jittering)

a. Extract crops

Pre-trained Network

CNN Feature Extractor

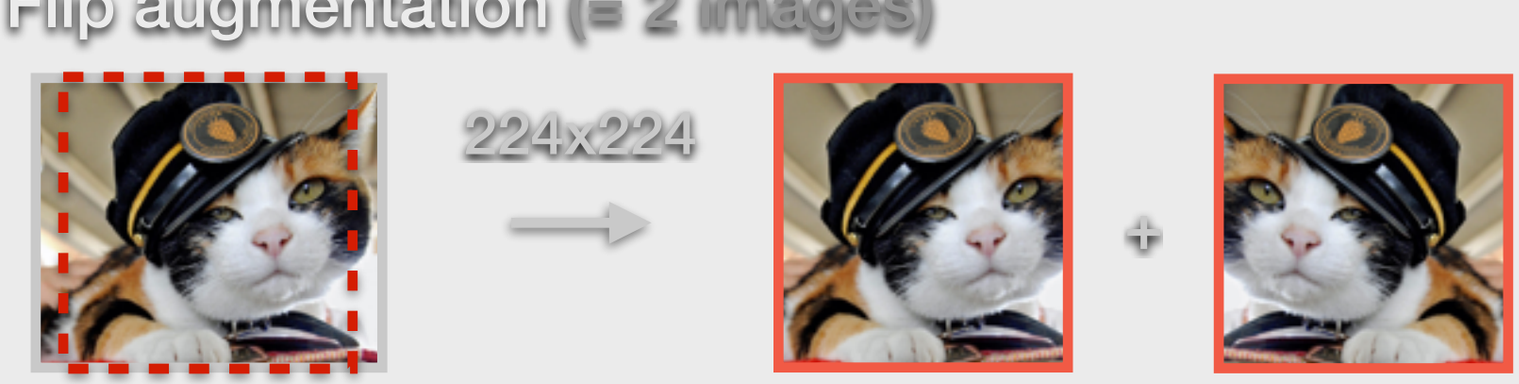


b. Pool features
(average, max)

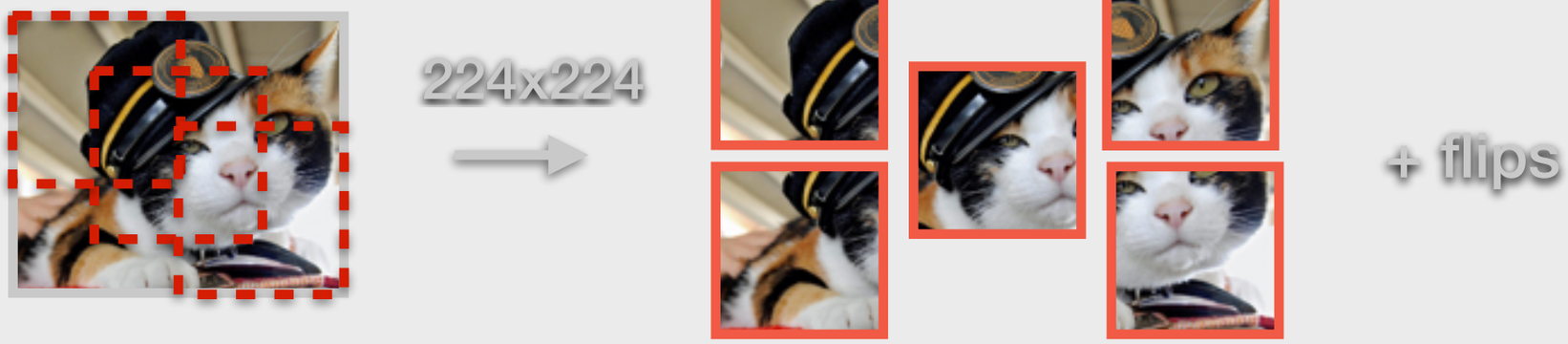
Data Augmentation

a. No augmentation (= 1 image)

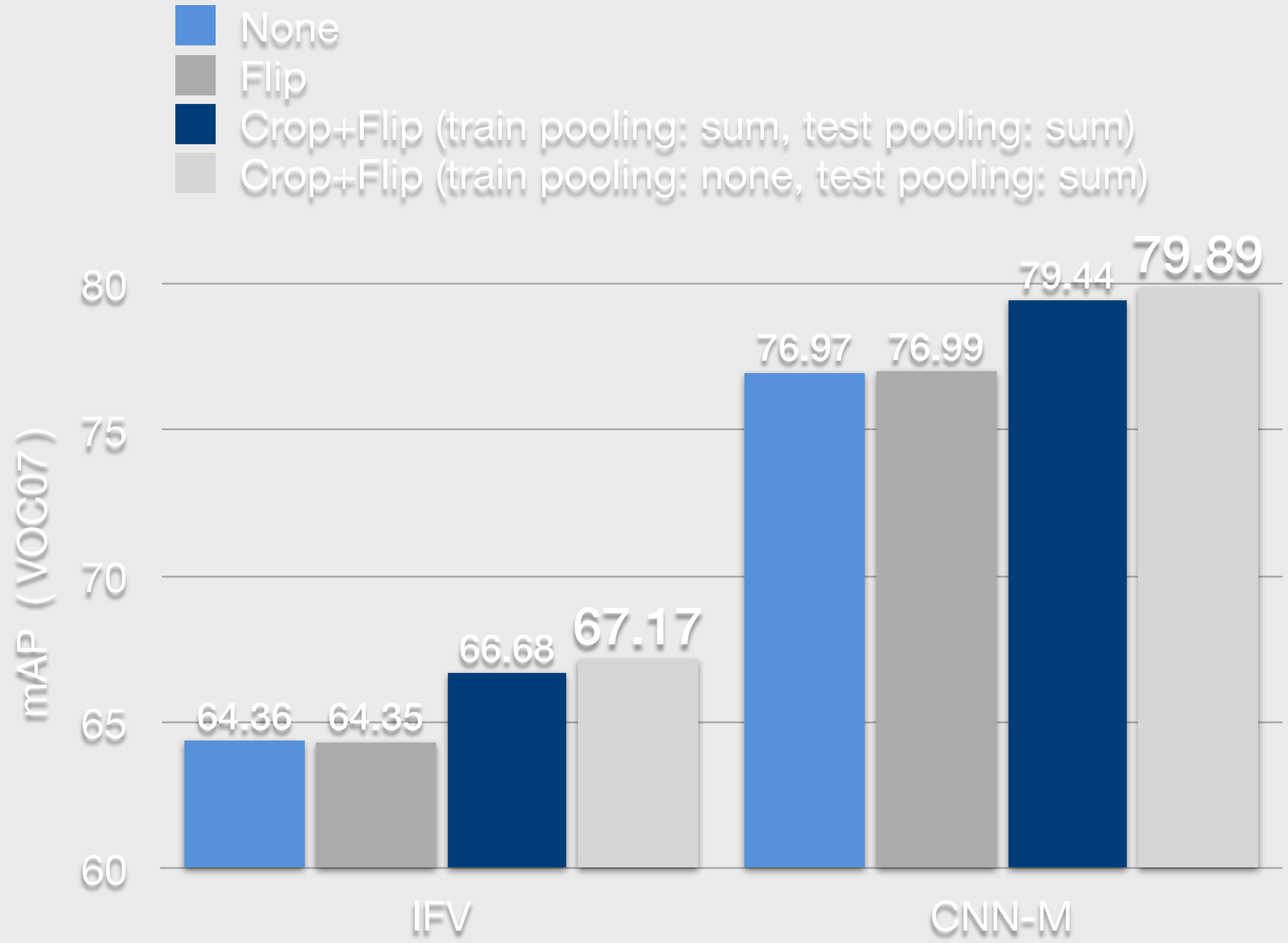
b. Flip augmentation (= 2 images)



c. Crop+Flip augmentation (= 10 images)



Data Augmentation



Outline

Study Introduction and Evaluation Setup

1 Different pre-trained networks

2 Data augmentation (for both CNN and IFV)

3 Dataset fine-tuning

4 Reducing CNN final layer output dimensionality

5 Colour and CNN / IFV

Fine-tuning

Network Pre-training
images from ILSVRC-2012

→

Pre-trained Network

→

General-purpose
Features

↓

Network Fine-tuning
images from target dataset

→

Fine-tuned Network

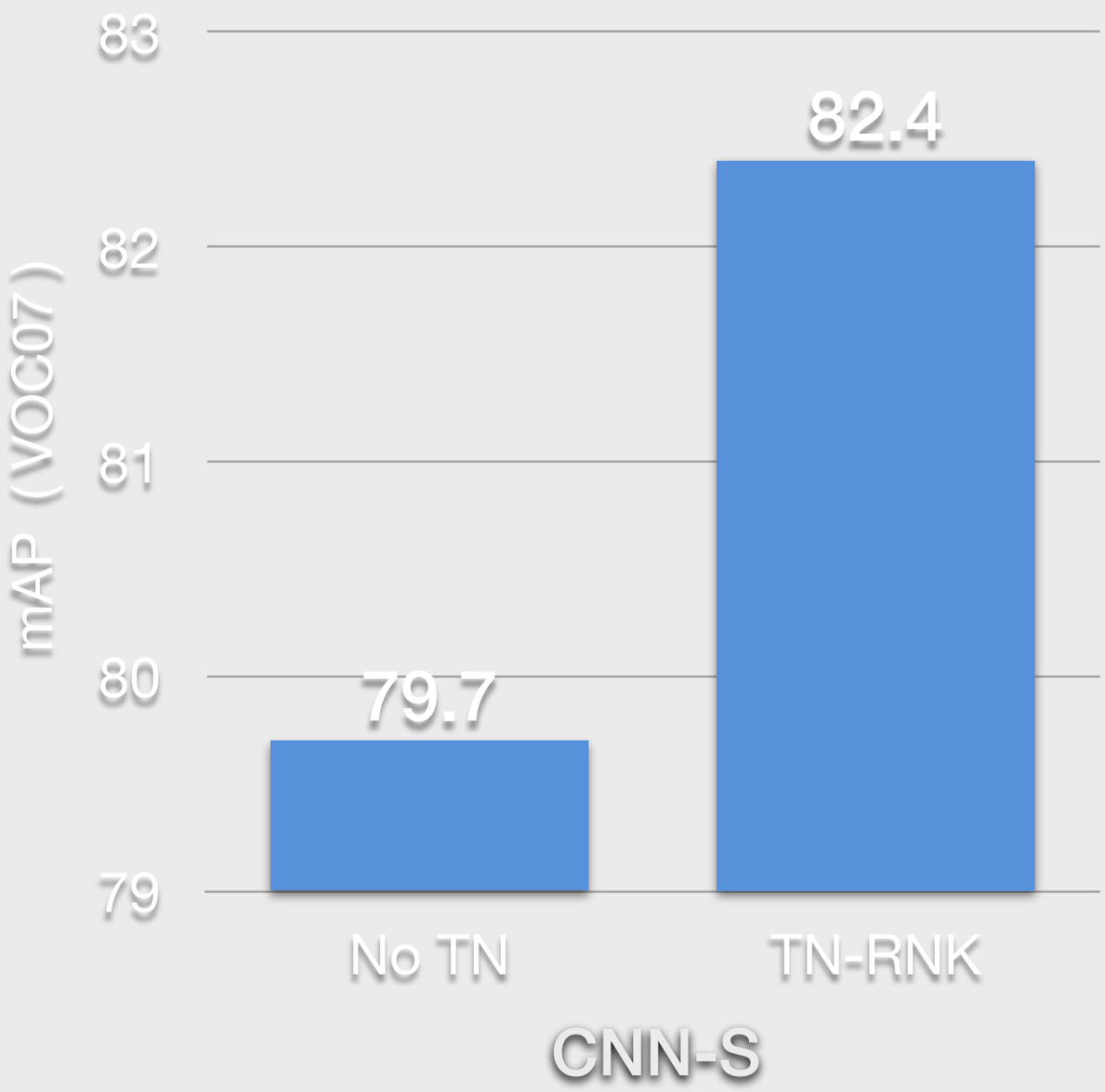
→

Dataset-specific
Features

For VOC 2007, the following loss functions were evaluated for the final fully connected layer:

- TN-CLS – classification loss $\max\{ 0, 1 - y w^T \phi(I) \}$
- TN-RNK – ranking loss $\max\{ 0, 1 - w^T(\phi(I_{POS}) - \phi(I_{NEG})) \}$

Fine-tuning



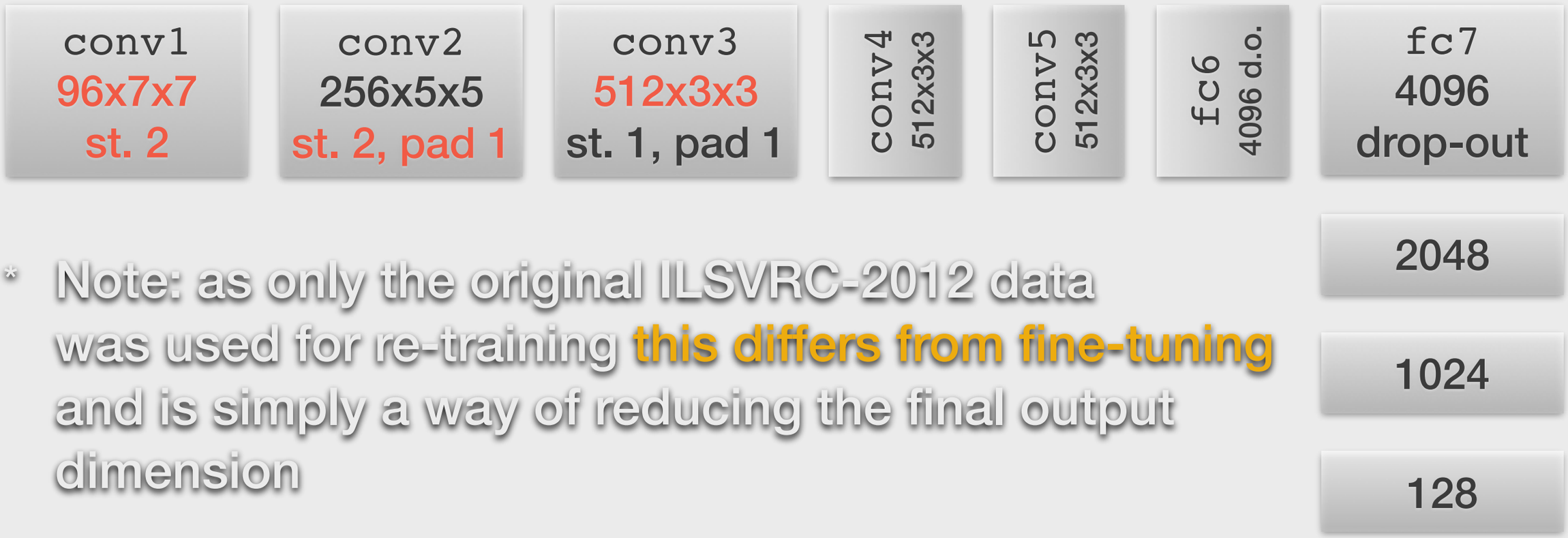
Outline

Study Introduction and Evaluation Setup

1. Different pre-trained networks
2. Data augmentation (for both CNN and IFV)
3. Dataset fine-tuning
4. Reducing CNN final layer output dimensionality
5. Colour and CNN / IFV

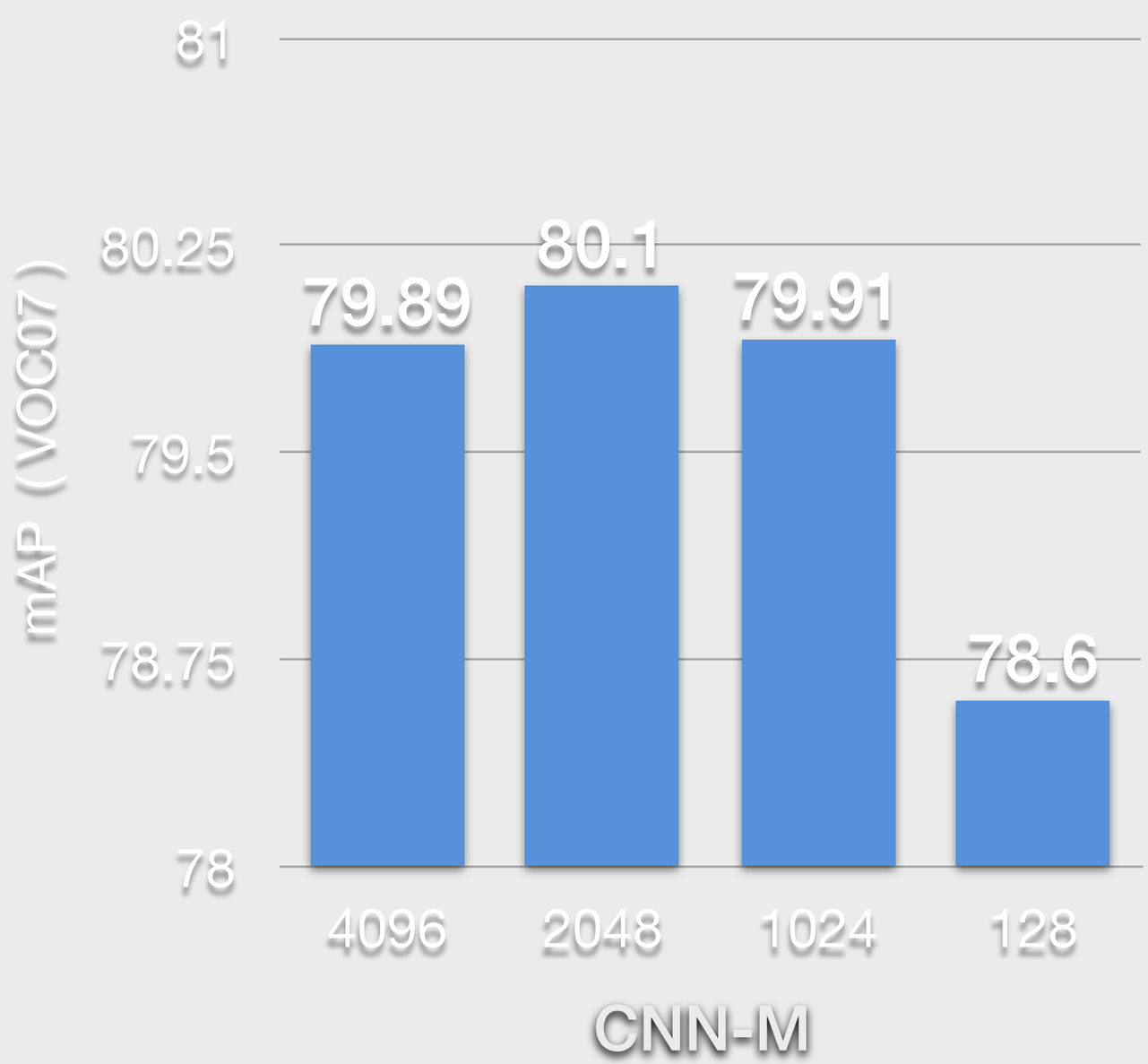
Low Dimensional CNN Features

- Baseline networks all have 4096-D last hidden layer
- We further trained three modifications to CNN-M with lower dimensional `full7` layers



* Note: as only the original ILSVRC-2012 data was used for re-training **this differs from fine-tuning** and is simply a way of reducing the final output dimension

Low Dimensional CNN Features

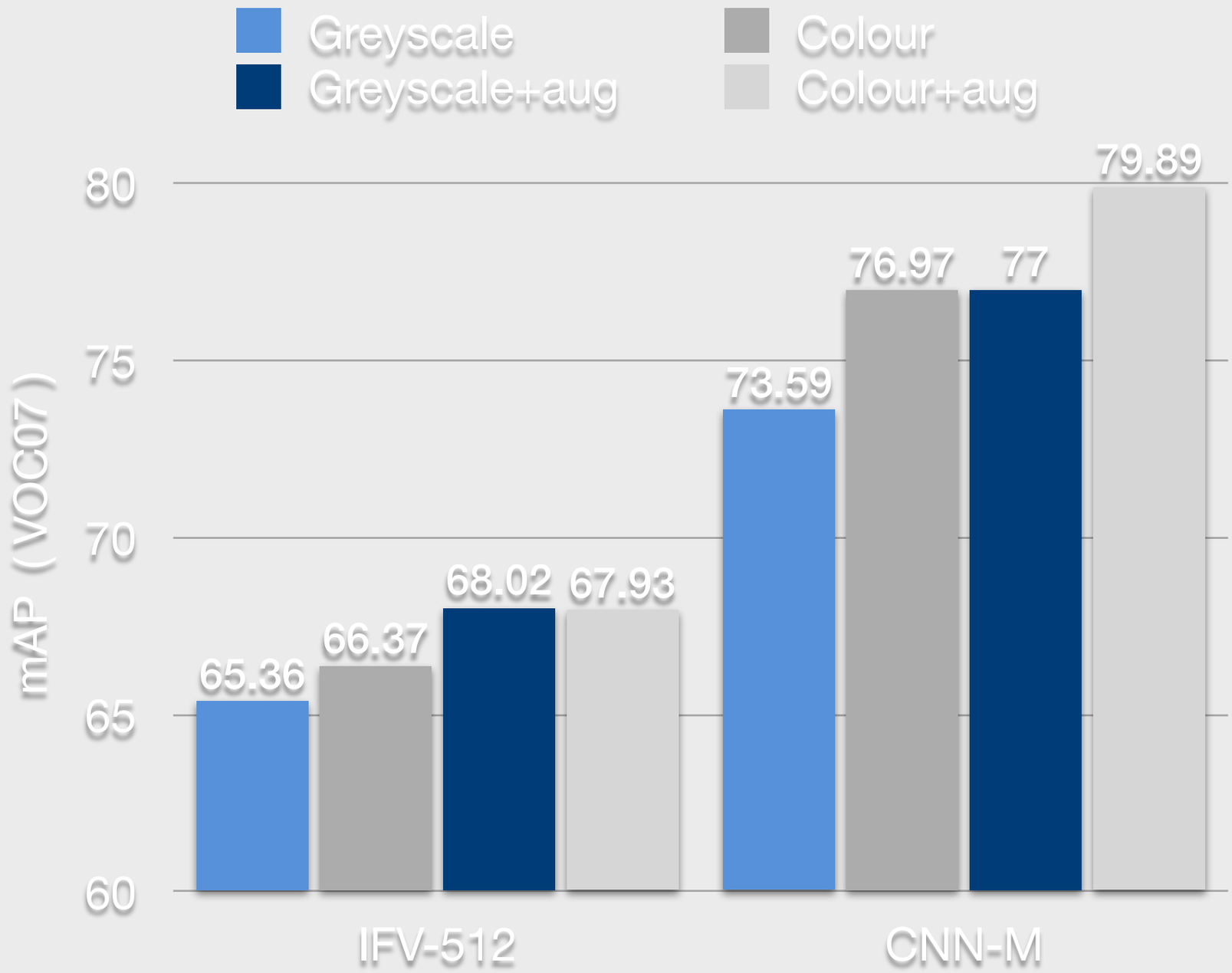


Outline

Study Introduction and Evaluation Setup

1. Different pre-trained networks
2. Data augmentation (for both CNN and IFV)
3. Dataset fine-tuning
4. Reducing CNN final layer output dimensionality
5. Colour and CNN / IFV

Impact of Colour



Comparison to State-of-the-art

	ILSVRC-2012	VOC2007	VOC2012
CNN-M 2048	13.5	80.1	82.4
CNN-S	**13.1**	79.7	82.9
CNN-S TUNE-RNK	**13.1**	**82.4**	**83.2**
Zeiler & Fergus	16.1		79.0
Oquab et al.	18.0	77.7	78.7 (82.8*)
Oquab et al.			***86.3****
Wei et al.		81.5 (***85.2****)	81.7 (***90.3****)

* Uses extended training data and/or fusion with other methods

Take Home Messages

- **CNN-based methods** >> shallow methods
- We can **transfer tricks** from deep features to shallow features
- We can achieve **incredibly low dimensional** (~128-D) but performant features with CNN-based methods
- If you get the details right, it's possible to get to state-of-the-art with **very simple methods**

There's more…

- Presented here was just a subset of the full results from the paper

- Check out the paper for full results on:
  - VOC 2007
  - VOC 2012
  - Caltech-101
  - Caltech-256
  - ILSVRC-2012

One more thing…

- CNN models and feature computation code can now be downloaded from the project website: http://www.robots.ox.ac.uk/~vgg/software/deep_eval/

- As before, **source code to reproduce all experiments will be made available**

Questions?