Synthetically-augmented data for deep text spotting

Andrea Vedaldi

VARVAI, ECCV 2016

http://adas.cvc.uab.es/varvai2016/

UNIVERSITY OF
OXFORD

Modern convolutional nets

[AlexNe… …. 2012]

Excellent **performance** in image understanding tasks

Learn a sequence of **general-purpose representations**

Millions of parameters learned from data

Generally **very large datasets** are required for good performance

The quest for **supervised** big data

The availability of **large annotated dataset** is perhaps the biggest limiting factor in machine learning applications. Challenges:

Data collection

- Cost (e.g. pictures of mars)
- Scarcity (e.g. rare diseases)
- Sensitivity (e.g. personal data, industrial secrets, military data)

Data annotation

- Scale (e.g. segmenting millions of images)
- Expertise (e.g. medical imaging)

A **major advantage of natural vision** is the ability to learn from only a few examples.

Ultimately, the cure are much smarter machines that require less supervision. But what can we do in the mean time?

Cheap sources of supervised data

Synthetic or augmented

Augmented data

or

Synthetic data

Synthetic data: generated using computer graphics.

Augmented data: generated by transforming real images.

Cheap sources of data

Synthetic or augmented

Augmented data

or

Synthetic data

Synthetic data: generated using computer graphics.

Augmented data: generated by transforming real images.

Augmented data

The standard approach is **data augmentation** / **jittering** / **virtual samples**

This amounts to apply random but semantic-preserving transformations to the image

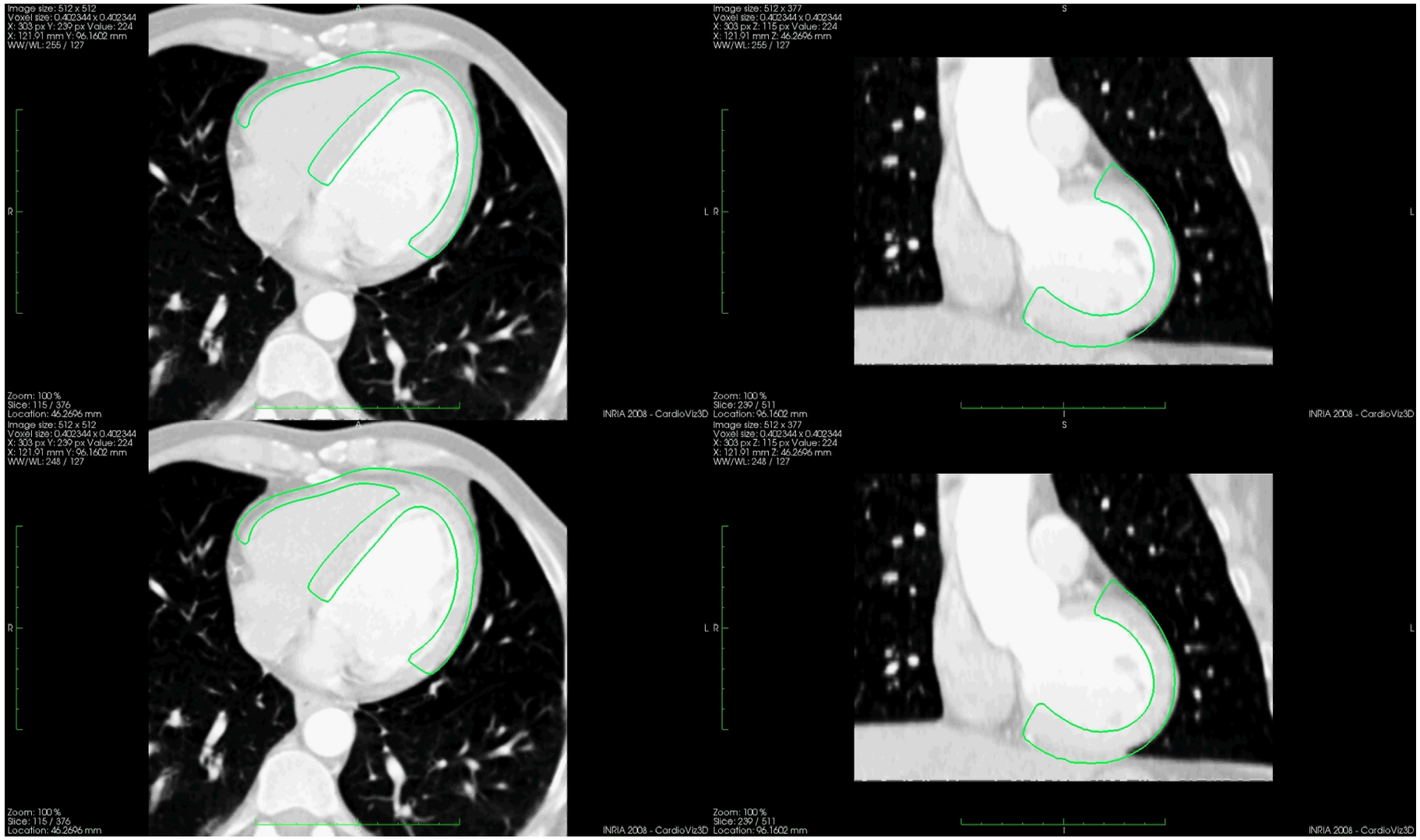
translation, scale changes rotation, affine distortions, colour shifts, saturation changes, noise, …

Here: data augmentation example from ResNet training

[S. Cho and K. Cha. Evolution of neural network training set through addition of virtual samples. In Proc. Evolutionary Computation, 1996]

Targeted transformations

E.g. to simulate motion in medical imaging



[A Prakosa, M Sermesant, P Allain, N Villain, C Rinaldi, K Rhode, R Razavi, H Delingette, N Ayache, Cardiac Electrophysiological Activation Pattern Estimation from Images using a Patient-Specific Database of Synthetic Image Sequences, IEEE Tr Biomedical Engineering 2013]

Cheap sources of supervised data

Synthetic or augmented

Augmented data

or

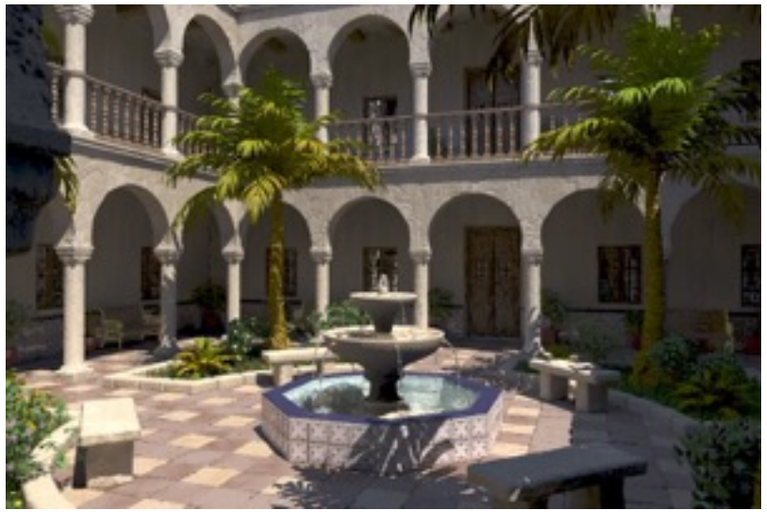
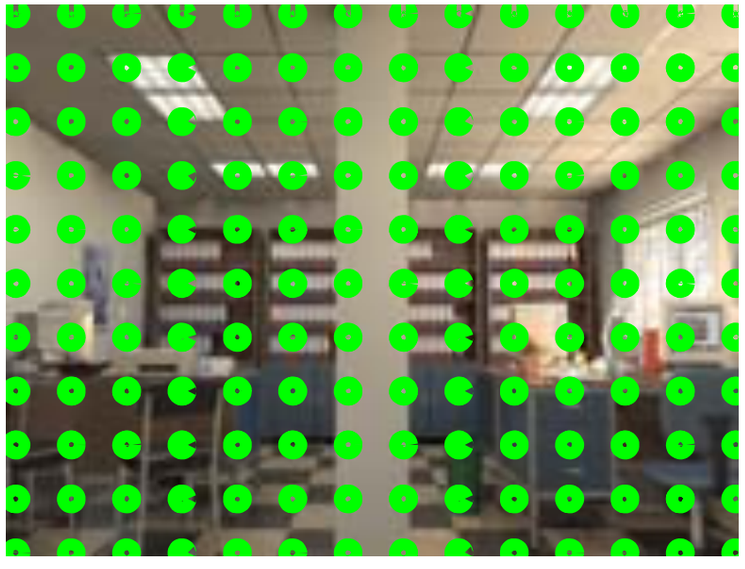
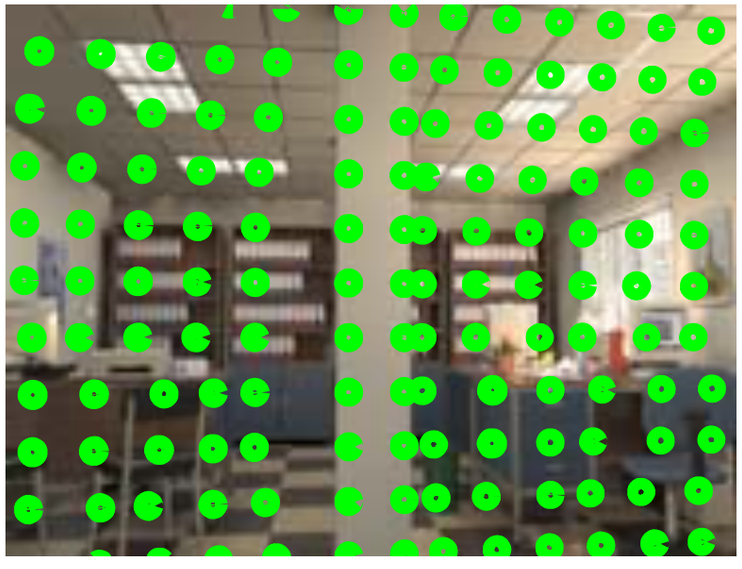
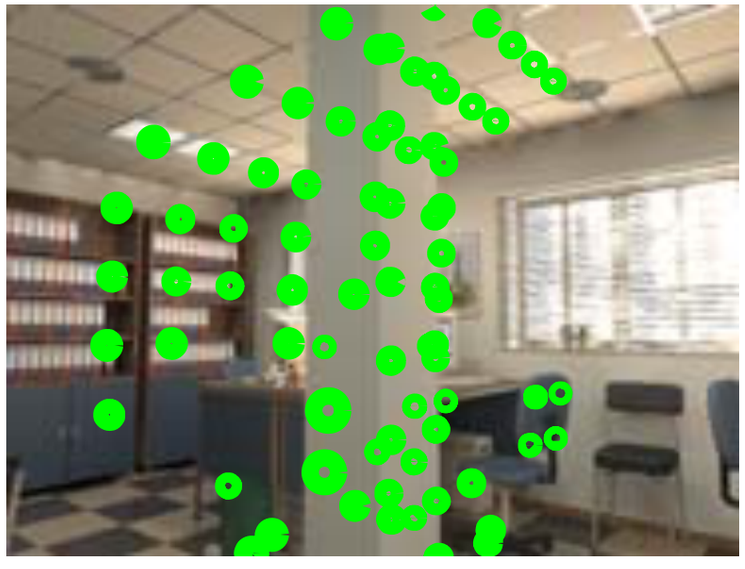
Synthetic data

Synthetic data: generated using computer graphics.

Augmented data: generated by transforming real images.

Synthetic data for learning local features



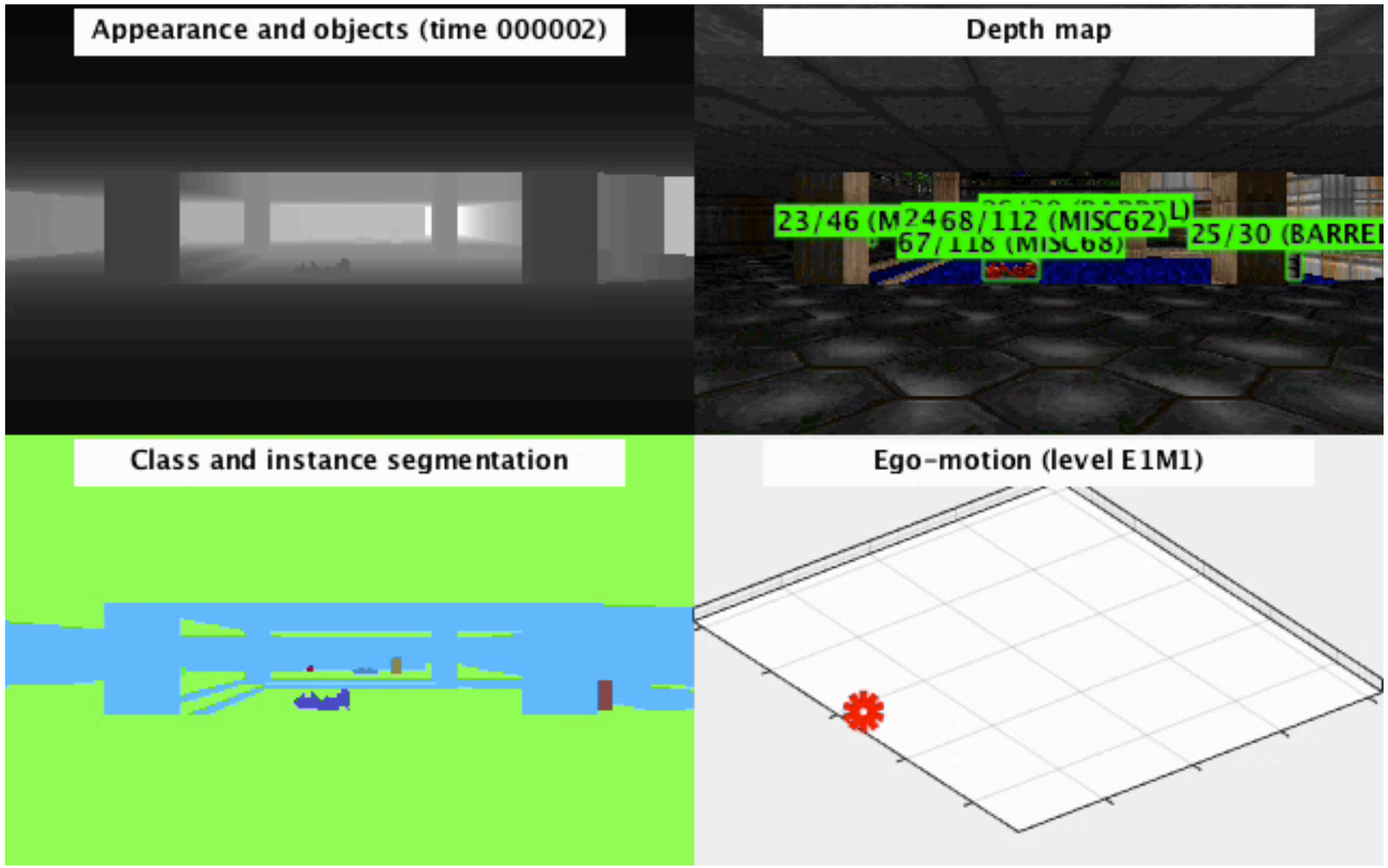
[A. Vedaldi, H. Ling, and S. Soatto. Knowing a good feature when you see it: Ground truth and methodology to evaluate local features for recognition. CVDRR, 2010]

Kinect

The best known example of synthetic training data

[Real-Time Human Pose Recognition in Parts from Single Depth Images, Shotton et al. 2016]

ResearchDoom



Pre-computed: Tons of data in **Microsoft Coco format**

Annotations: Object masks (instance and categories), depth maps, egomotion

http://www.robots.ox.ac.uk/~vgg/research/researchdoom/

Some references

Games

- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing Atari with deep reinforcement learning. CoRR, 2013.
- Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action- conditional video prediction using deep networks in Atari games, NIPS, 2015.
- W. Qiu and A. Yuille. UnrealCV: Connecting computer vision to Unreal Engine. arXiv, 2016.
- M. Kempka, M. Wydmuch, G. Runc, J. Toczek, and W. Jaskowski. ViZDoom: A doom-based AI research platform for visual reinforcement learning. CoRR, abs/1605.02097, 2016.
- S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In Proc. ECCV, 2016.

Optical Flow

- Sintel: D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In Proc. ECCV, 2012.
- Flying Things etc: N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proc. CVPR, 2016.

Other simulations

- C. Chen, A. Seff, A. Kornhauser, and J. Xiao. DeepDriving: Learning affordance for direct perception in autonomous driving. In Proc. ICCV, 2015.
- G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In Proc. CVPR, 2016.

Cheap sources of data

Synthetic or augmented

Augmented data

or

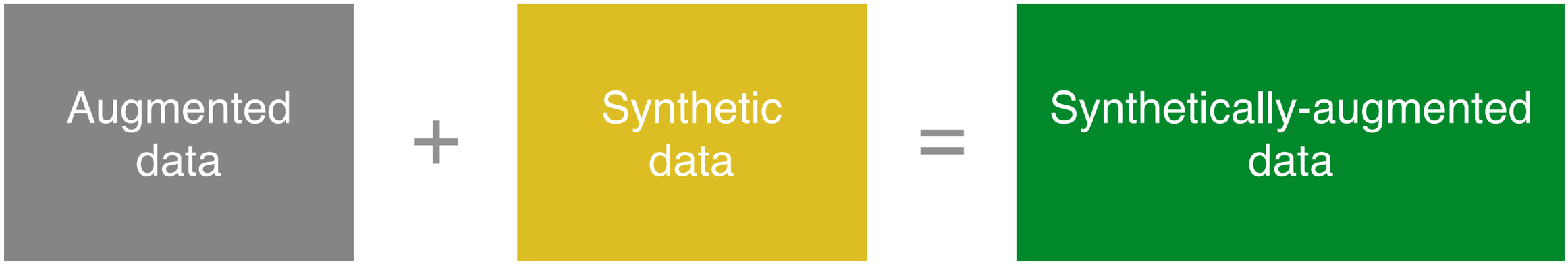
Synthetic data

Synthetic data: generated using computer graphics.

Augmented data: generated by transforming real images.

Cheap sources of data

Synthetic or augmented



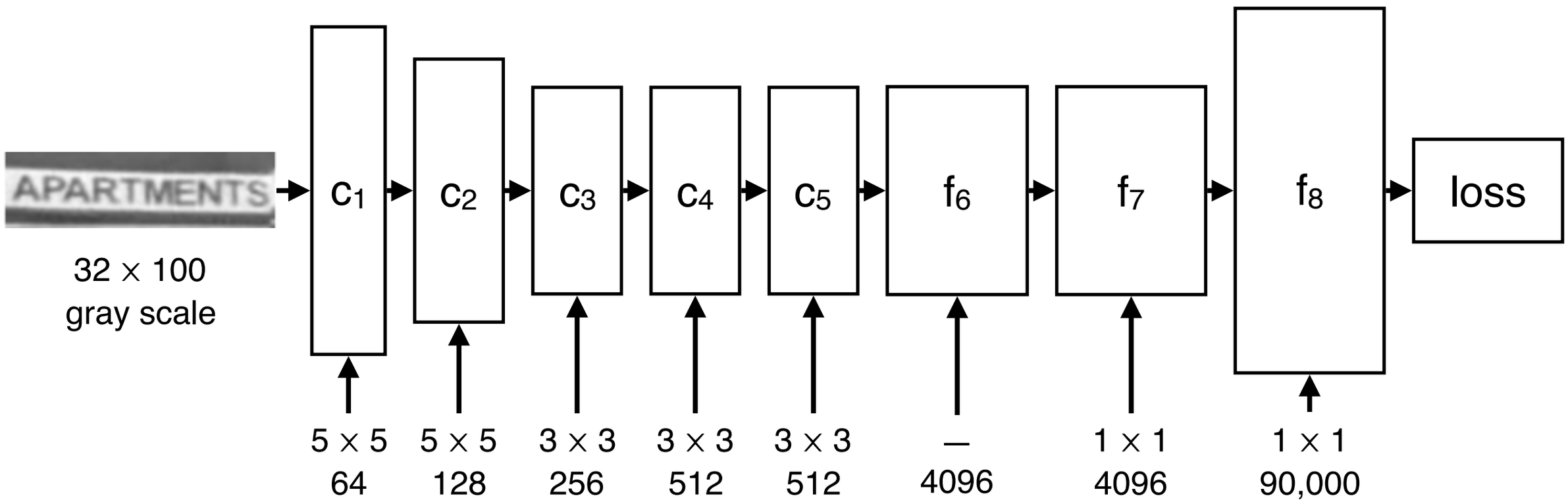
Modify **real images** by inserting **virtual objects.**

Synthetic data for text recognition

Synthetically-augmented data for text detection

Synthetic data mixing for face/scene retrieval

A massive classifier



Goal: map images to one of 90K classes (one per word)

Architecture

- each linear operator is followed by ReLU
- c_1, c_2, c_3, c_5 are followed by 2×2 max pooling
- 500 million parameters
- evaluation requires 2.2ms on a GPU

Learning a massive classifier

Massive training data

- ~100 examples per word
- 9 million images for 90K words

Learning algorithm

- SGD
- mini-batches

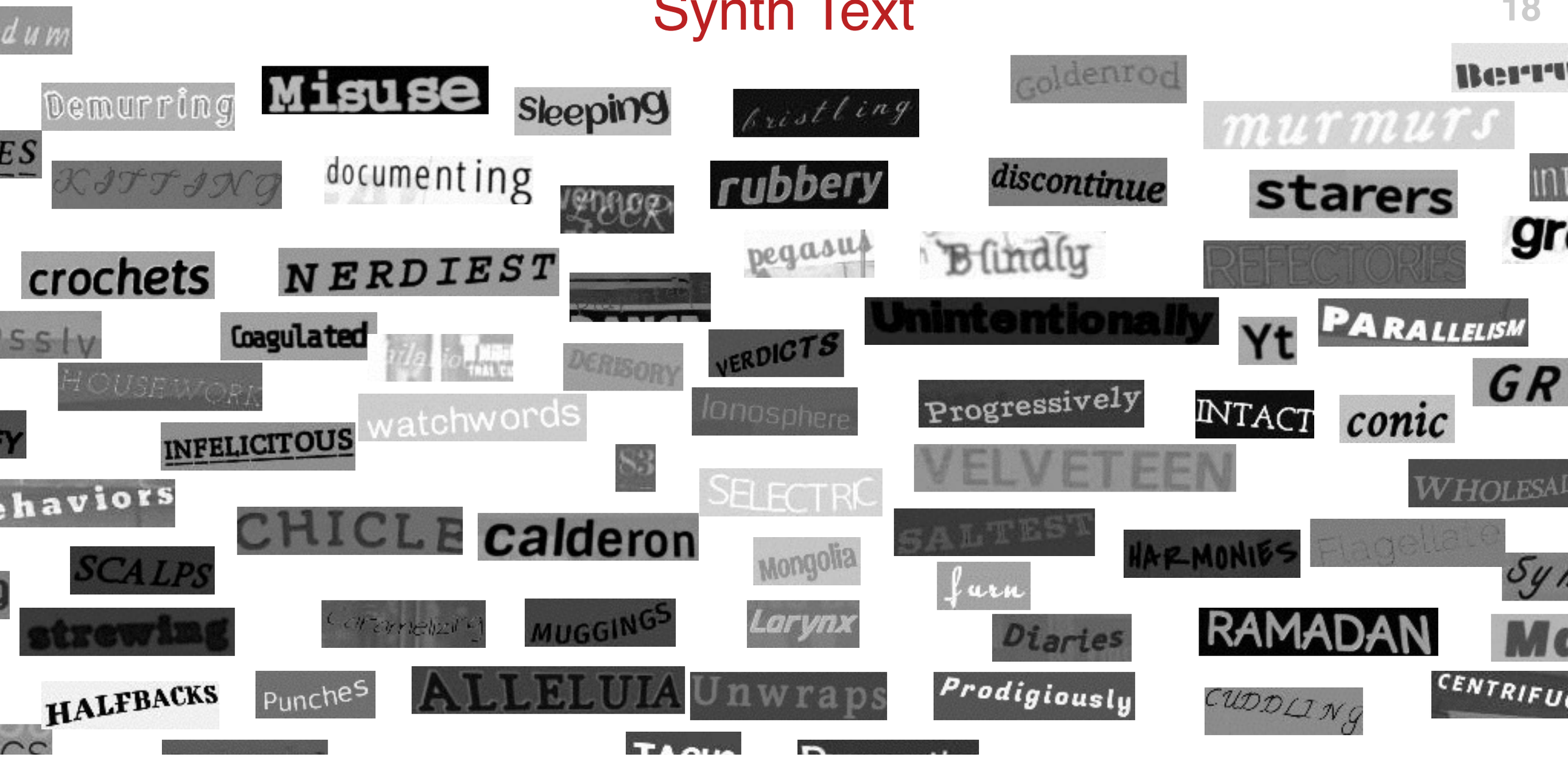
Mini-batch composition

- stable learning requires each batch to contain ~1/5 of all the classes
- batch size = 18K (too slow!)

Incremental training

- learn first using 5K classes only (batch size = 1K)
- then incrementally add 5K more classes

Synth Text

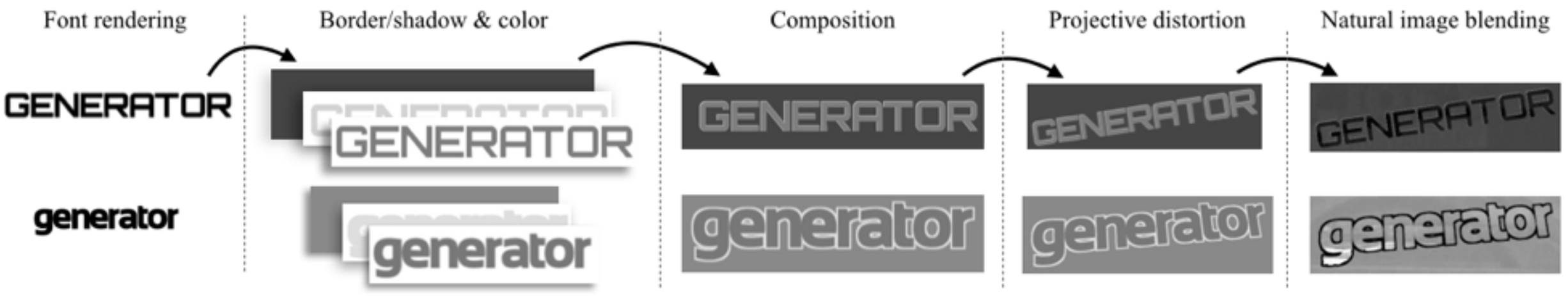


infinity large, fully supervised (at the word and character level)

http://www.robots.ox.ac.uk/~vgg/data/text/

9M precomputed images [10 GB]

Synth Text generation



Font rendering

- sample at random one of 1400 Google Fonts

Border/shadow

- randomly add inset/outset border and shadow

Projective distortion

Blending

- use a random crop from SVT as background
- randomly sample alpha channel, mixing operator (normal, burn, …)

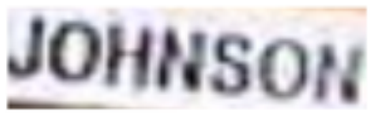
Noise

- elastic distortion, white noise, blur, JPEG compression, …

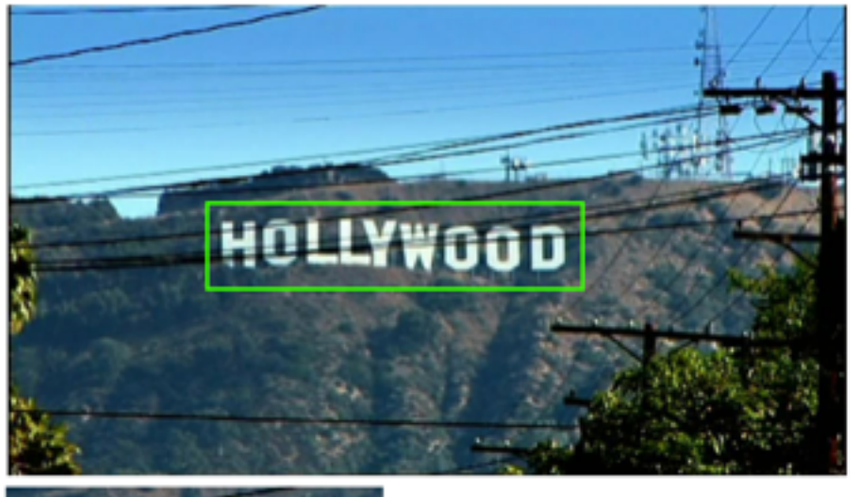
Qualitative results: text retrieval

“APARTMENTS”

“BORIS JOHNSON”



“HOLLYWOOD”

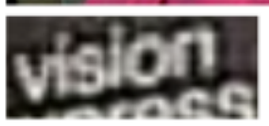


Qualitative results: text retrieval

“POLICE”

“CASTROL”

“VISION”



Demo

http://zeus.robots.ox.ac.uk/textsearch/#/search/

Synthetic data for text recognition

Synthetically-augmented data for text detection

Synthetic data mixing for face/scene retrieval

Text detection

Recognition is only half of the problem: text needs to be detected in the first place

Goal: Train a fast, fully-convolutional ConvNet for localisation
on this synthetic data

Key challenge: availability of suitable training data

Text spotting in natural scenes is hard

Nuisance factors

Fonts

Distortions

Colors

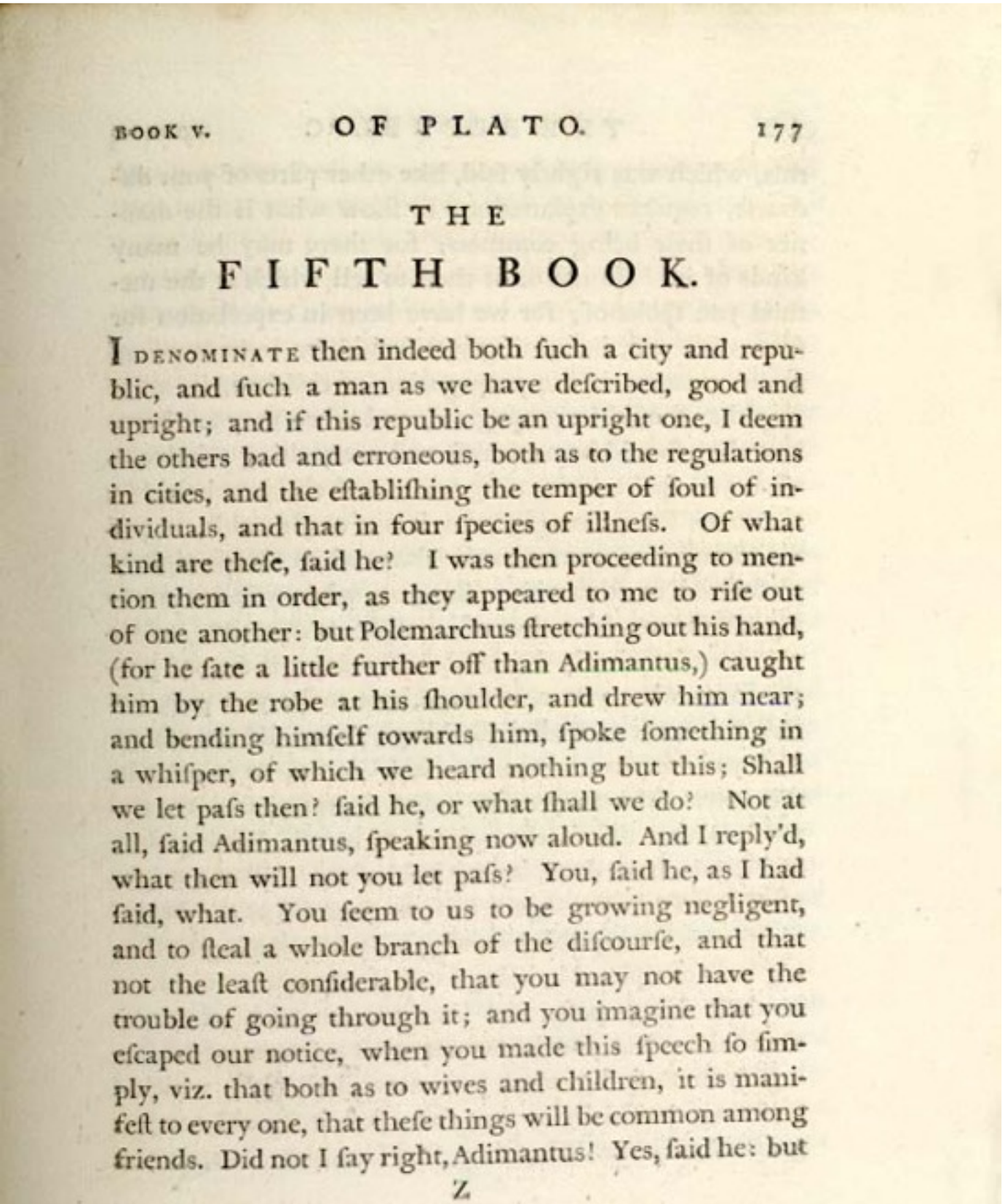
Blur

Shadows

Borders

Textures

Sizes …



THE

FIFTH BOOK.

I DENOMINATE then indeed both ſuch a city and republic, and ſuch a man as we have deſcribed, good and upright; and if this republic be an upright one, I deem the others bad and erroneous, both as to the regulations in cities, and the eſtabliſhing the temper of ſoul of individuals, and that in four ſpecies of illneſs. Of what kind are theſe, ſaid he? I was then proceeding to mention them in order, as they appeared to me to riſe out of one another: but Polemarchus ſtretching out his hand, (for he ſate a little further off than Adimantus,) caught him by the robe at his ſhoulder, and drew him near; and bending himſelf towards him, ſpoke ſomething in a whiſper, of which we heard nothing but this; Shall we let paſs then? ſaid he, or what ſhall we do? Not at all, ſaid Adimantus, ſpeaking now aloud. And I reply'd, what then will not you let paſs? You, ſaid he, as I had ſaid, what. You ſeem to us to be growing negligent, and to ſteal a whole branch of the diſcourſe, and that not the leaſt conſiderable, that you may not have the trouble of going through it; and you imagine that you eſcaped our notice, when you made this ſpeech ſo ſimply, viz. that both as to wives and children, it is manifeſt to every one, that theſe things will be common among friends. Did not I ſay right, Adimantus! Yes, ſaid he: but

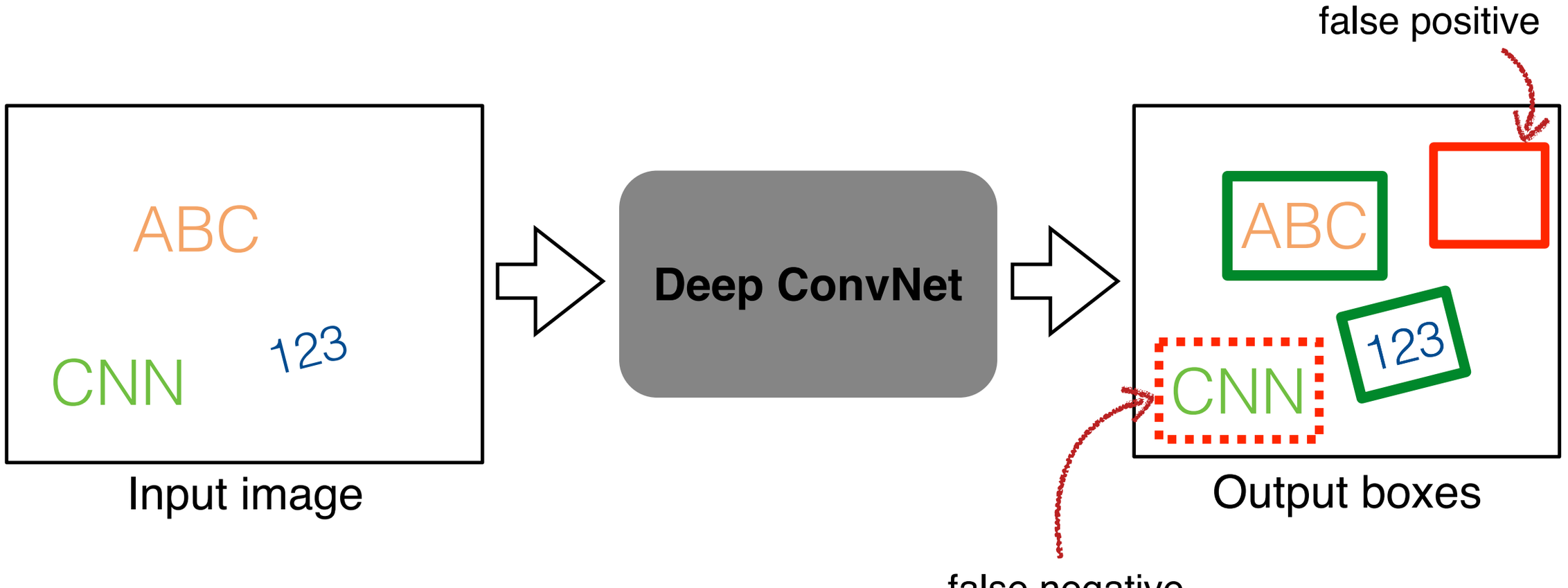
High Contrast

Plain Background

Well defined lines

Limited variation in fonts and size

Learning a CNN for detection



Learning a detection CNN requires thousands of images

Supervised training requires to know where all the text instances are

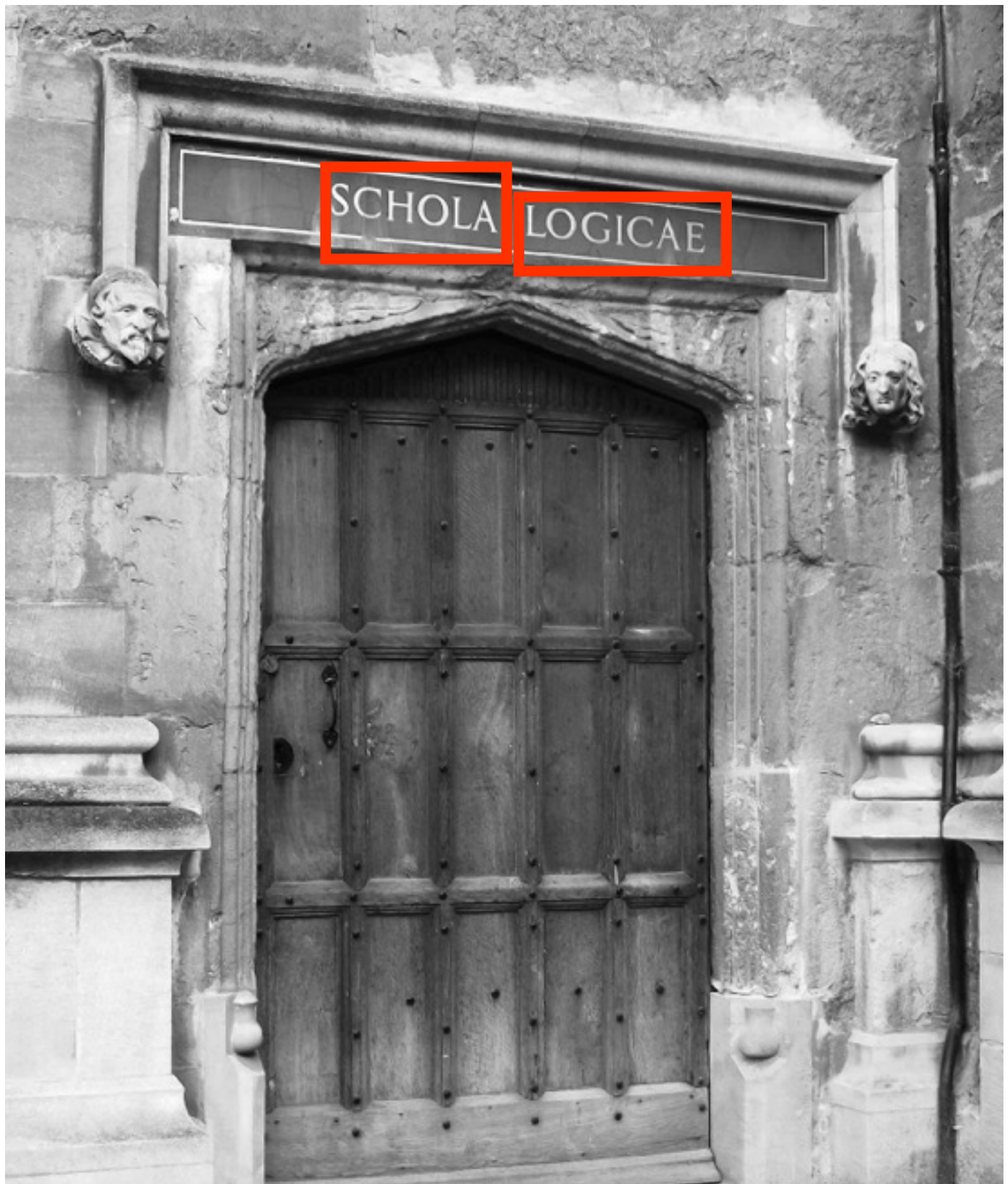
Only in this way we can tell correct detections from false positives (incorrect ones) and false negatives (missing ones)

Generating realistic text in scenes

Much more challenging than generating only text

Easy

Difficult



Synth Scene Text

Synthetically-augmented real data

Input Image

Synthetic Scene Text Image

A fully-automatic and fast procedure (0.5sec/image)
aware of the 3D geometry of the scene

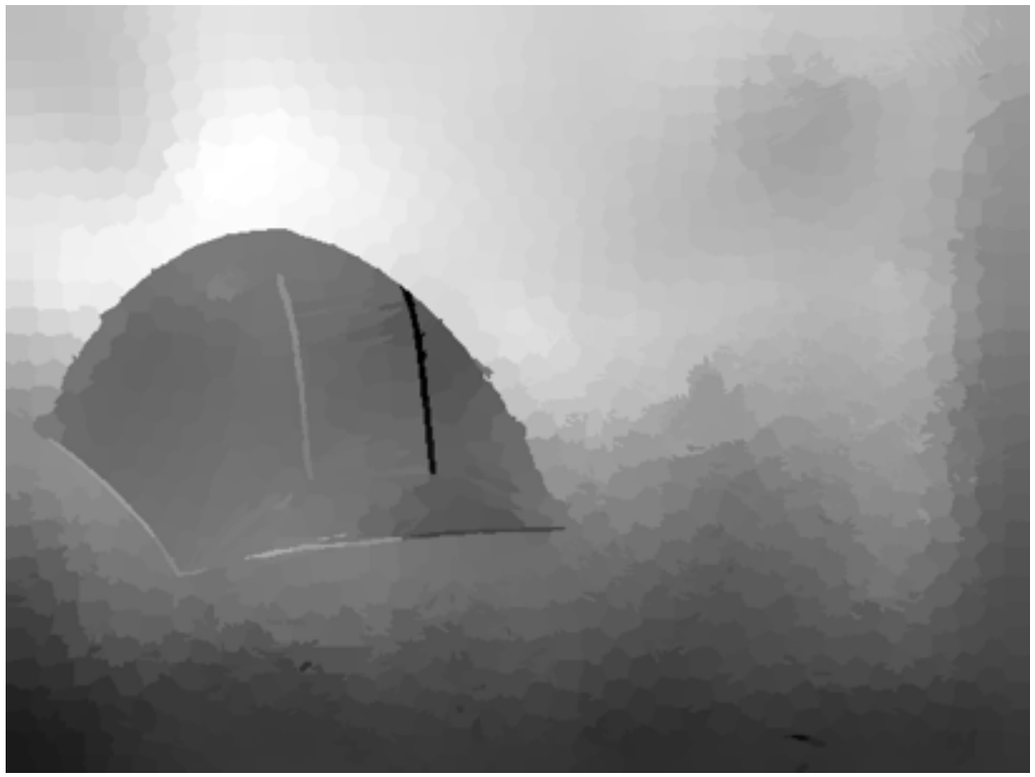
Word and character level annotations possible

800K pre-computed images available for download
http://www.robots.ox.ac.uk/~vgg/data/scenetext/

Synth Scene Text pipeline

Step 1: Predict the depth image

Input Image



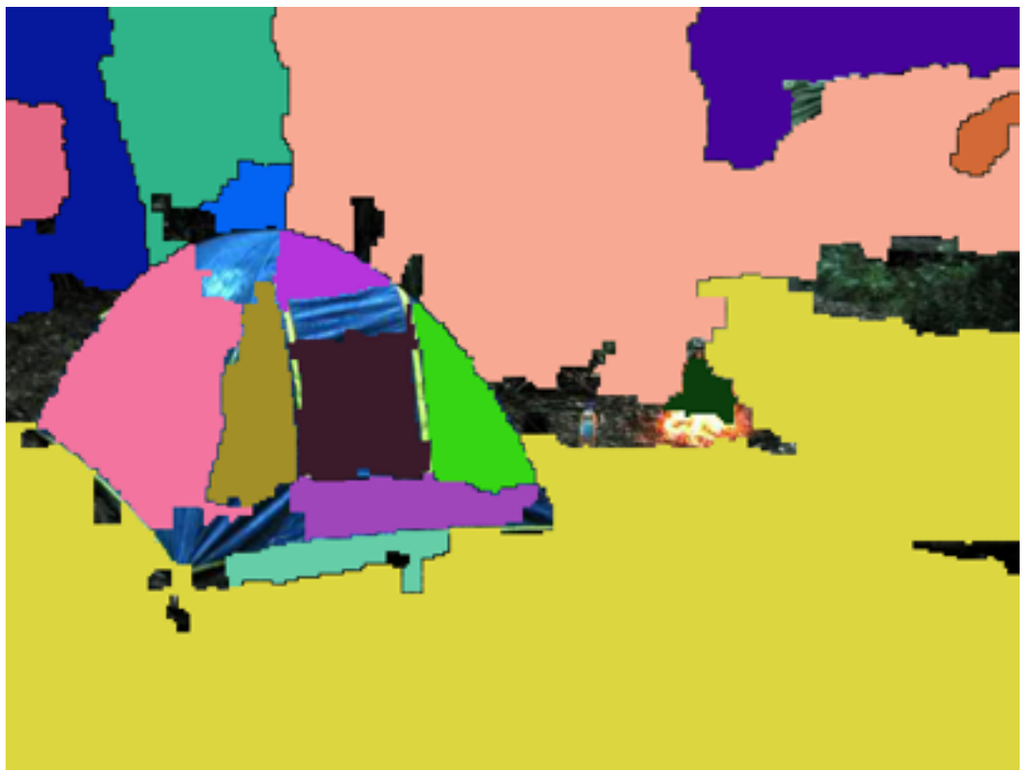
Depth Image

Monocular depth prediction uses the CNN by [Liu et al., CVPR 2015]

Synth Scene Text pipeline

Step 2: Find homogeneous surfaces

Input Image



Surfaces

Find regions that are likely to belong to the same 3D surface

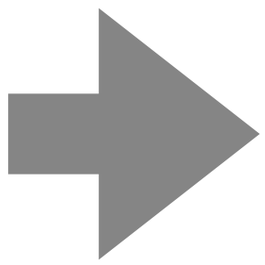
Goal: avoid straddling occlusion boundaries

Segmentation uses the gPb-UCM regions [Arbelaez et al. PAMI 2011]

Synth Scene Text pipeline

Step 3: Render scene text

Input Image



Virtual scene text

Place the text respecting local geometry and boundaries

Generate text styles as for Synth Text

Blend using Poisson Composition [Perez et al., TOG 2003]

Geometry-aware synthetic text

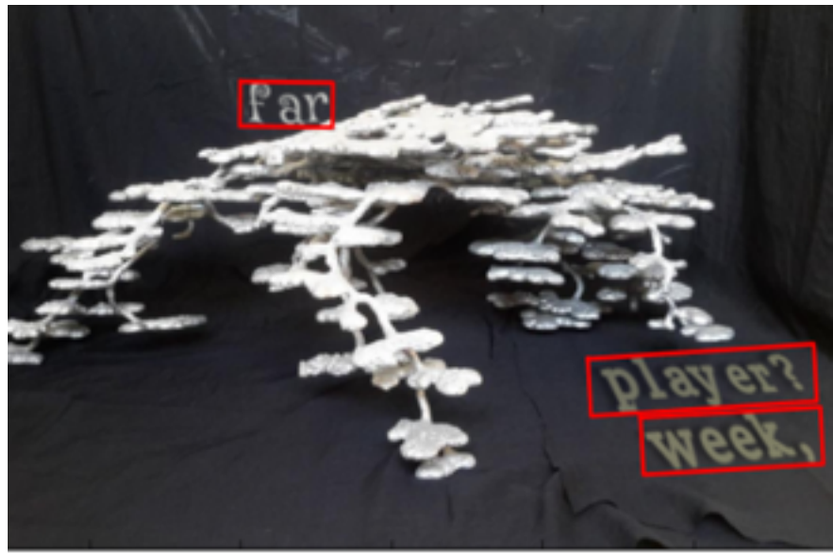

Apr
Mikes
and
Mes
agree


motion


Very
We
for
Fine
LAYERS
coin
The


buy
FOR
BOS,


From:
the
like
Get
a
the
and
is
are


far
player?
week,


the
can
you
then
disk


colors)


wire.
wire
the
allowed to happen...


and
the
(In


SOME


one

Comparison with existing datasets

ICDAR 2013

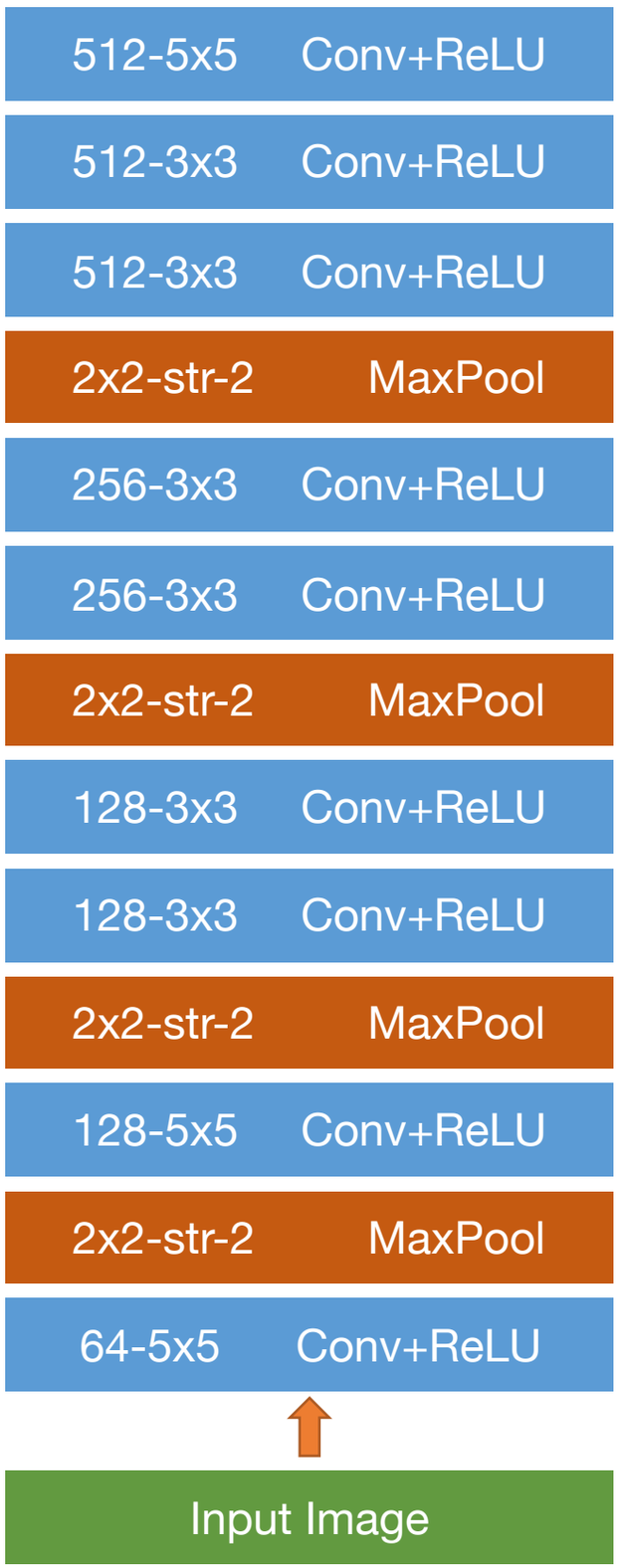
[Karatzas et al., ICDAR 2013]

Street View Text (SVT)

[Wang et al., ECCV 2010]

Datasets	Number of Images		Number of Words	
	Train	Test	Train	Test
ICDAR{11,13,15}	229	255	849	1095
SVT	100	249	257	647
Ours	**858,750**		**7,266,866**	

A fast CNN for text detection



Our CNN architecture d**ensely regresses text bounding boxes**

It combines ideas from

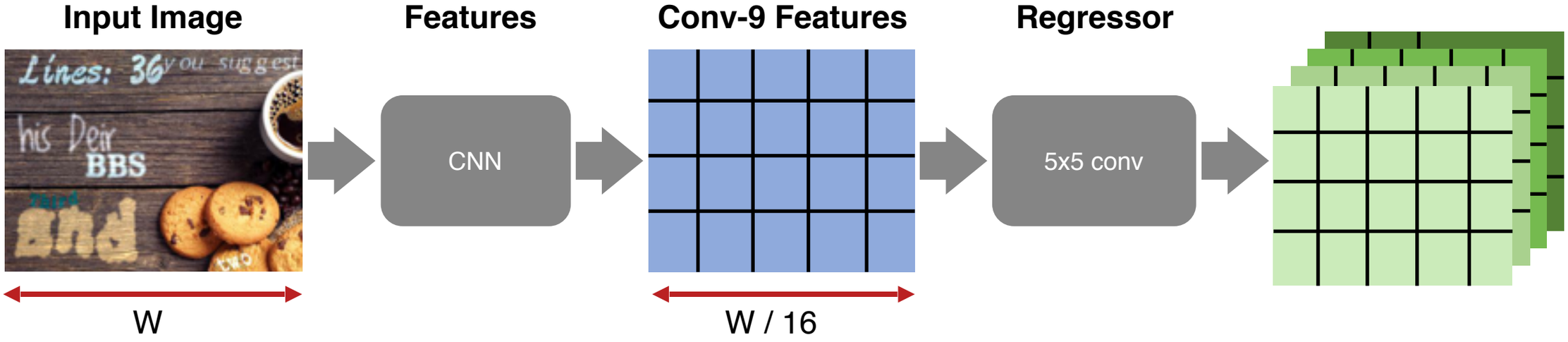
- You Only Look Once (YOLO) [Redmon et al., CVPR 2016]
- Fully Convolutional Network [Long et al., CVPR 2015]

Modifications to YOLO for text detection:

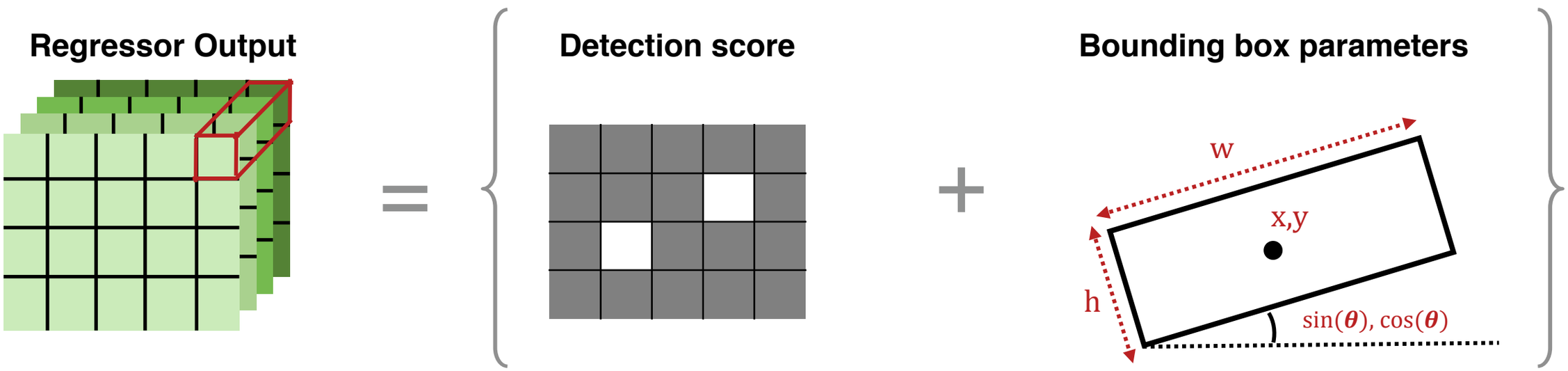
- Denser output for small text instances
- Fully convolutional for high resolution images

Convolutional YOLO

The CNN extract features with a stride of 16 pixels; then a linear convolutional regressor predicts one bounding box per quantised location



The regressor predicts 7 parameters: the detection score and the box geometry



Evaluation

Results

Text Localisation

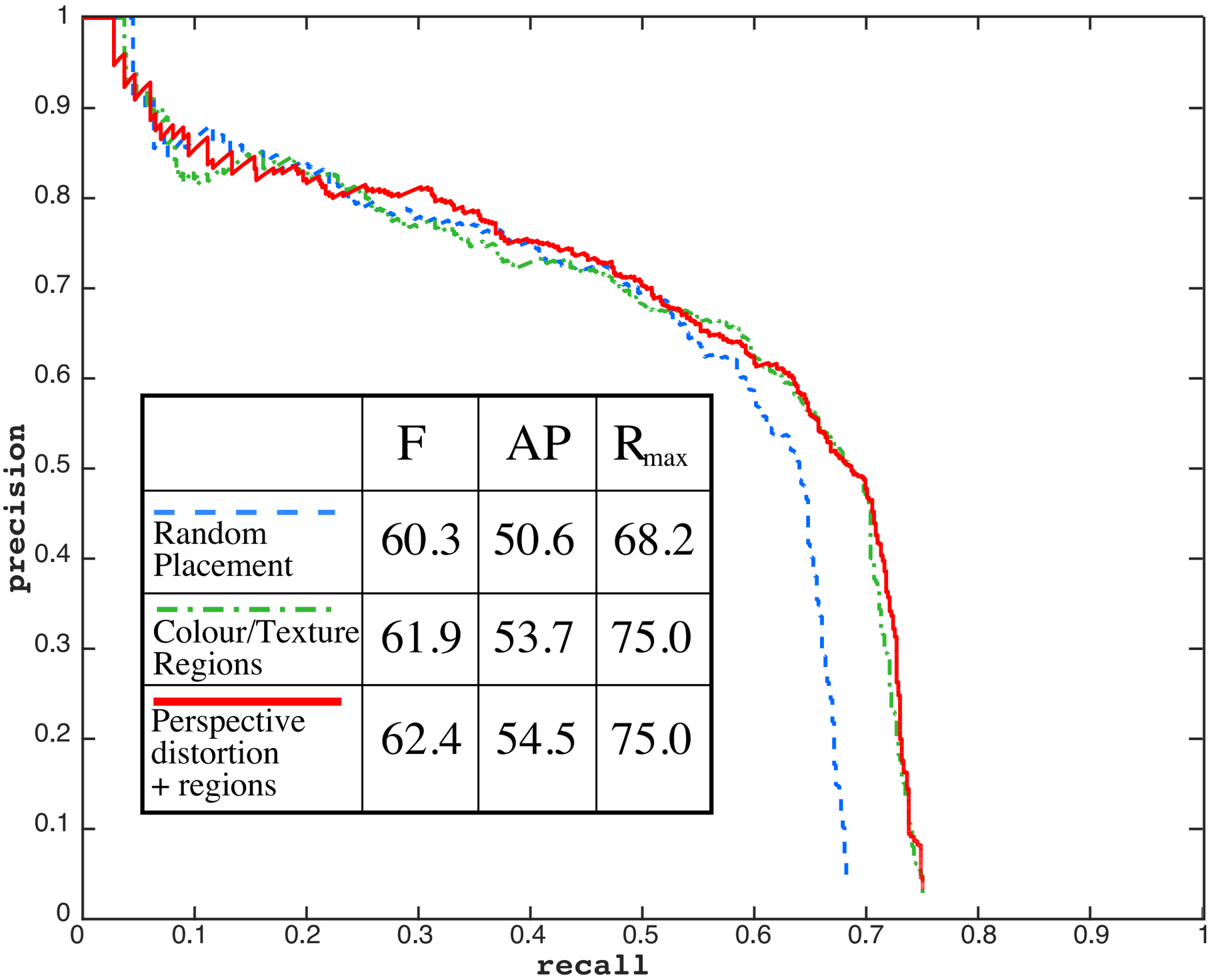
		ICDAR 11	ICDAR 13	SVT
Neumann	ICCV 13	72.3	-	-
Jaderberg	IJCV 15	76.8	76.8	24.7
Huang	ECCV 15	78	-	-
Zhang	CVPR 15	80	80	-
Ours		82.3	83.0	26.7

End-to-End Text Spotting

		ICDAR 11	ICDAR 13	SVT
Neumann	ICCV 13	45.2	-	-
Jaderberg	IJCV 15	69	76	53
Ours		81.0	84.7	55.7

Importance of realism

Benefit of increasingly complex text generation



	F	AP	R_{max}
Random Placement	60.3	50.6	68.2
Colour/Texture Regions	61.9	53.7	75.0
Perspective distortion + regions	62.4	54.5	75.0

EXIT 28B

Brooklyn Br

RIGHT LANE EXIT ONLY

11' - 0" CLEARANCE
NO COMMERCIAL TRAFFIC

UNLAWFUL
TO CROSS
SOLID LINE
ON
OR SIDE

WATER ST

EVERIT ST


EXIT 28B
Brooklyn Br
RIGHT LANE EXIT ONLY
11'-0" CLEARANCE
NO COMMERCIAL TRAFFIC
WATER ST
EVERIT ST

BANANA CREAM PIE

BUTTER AND TOAST

BLACK FOREST CAKE

Cinnamon Sugar

LEMON MERINGUE

COOKIES AND CREAM

CHOCOLATE PISTACHIO

GRANOLA & YOGURT

BEER PRETZEL

LEMON RICOTTA DOUGHNUT HOLES WITH OLIVE OIL GLAZE

ELVIS AND MARSHMALLOW

MAPLE

This is not a nut free kitchen
Please let us know

ORGANIC FAIR TRADE COFFEE

1.50 + tax

BANANA CREAM PIE

BUTTER AND TOAST

BLACK FOREST CAKE

Cinnamon Sugar

LEMON MERINGUE

COOKIES AND CREAM

CHOCOLATE PISTACHIO

GRANOLA & YOGURT

BEER

PRETZEL

LEMON RICOTTA DOUGHNUT HOLES WITH OLIVE OIL GLAZE

ELVIS AND MARSHMALLOW

MAPLE

This is not a nut free kitchen Please let us know

ORGANIC FAIR TRADE

COFFEE

150
1 shot

HOUSTON

INSPIRED
HIP TASTY
FUNKY SAVVY


HOUSTON
INSPIRED
HIP TASTY
FUNKY SAVVY

1998
20 NOV
15:55
DYNRNG
47 db
35CV76
DEPTH
138 MM
POWER
40%
FPS
26
REJECT
1
EDGE
2
GREY
3
SMOOTH
2
ALPHA

CINELOOP(R) 32

0 mm

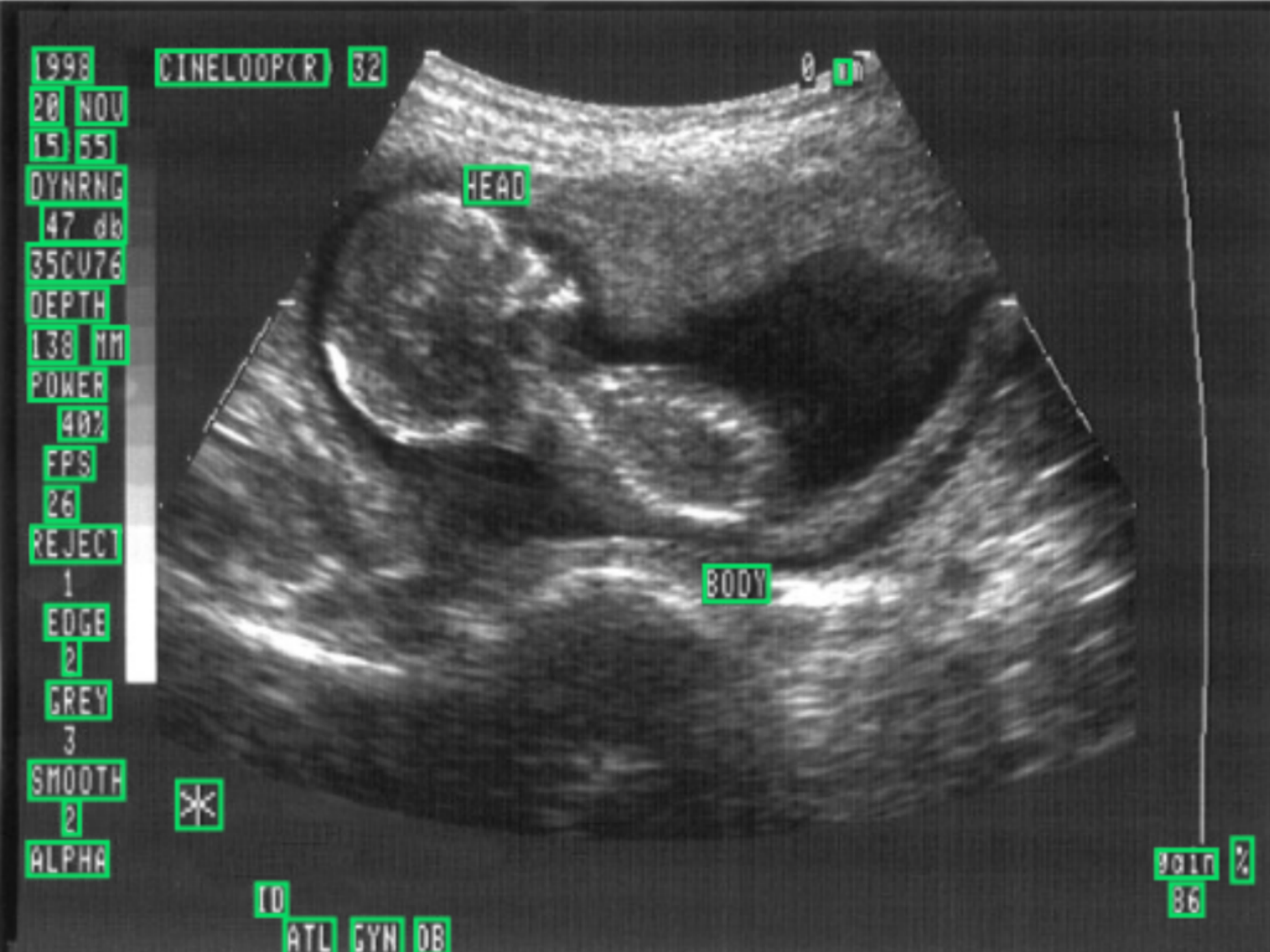
HEAD

BODY

gain %
86

ID

ATL GYN OB


1998
20 NOV
15 55
DYNRNG
47 db
35CV76
DEPTH
138 MM
POWER
40%
FPS
26
REJECT
1
EDGE
2
GREY
3
SMOOTH
2
ALPHA
CINELOOP(R) 32
HEAD
BODY
0
1
GAIN
86
CD
ATL GYN OB

Demo

zeus.robots.ox.ac.uk/textspot/

Synthetic data for text recognition

Synthetically-augmented data for text detection

Synthetic data mixing for face/scene retrieval

Compound query retrieval

Search for **specific people** in **specific scenes**

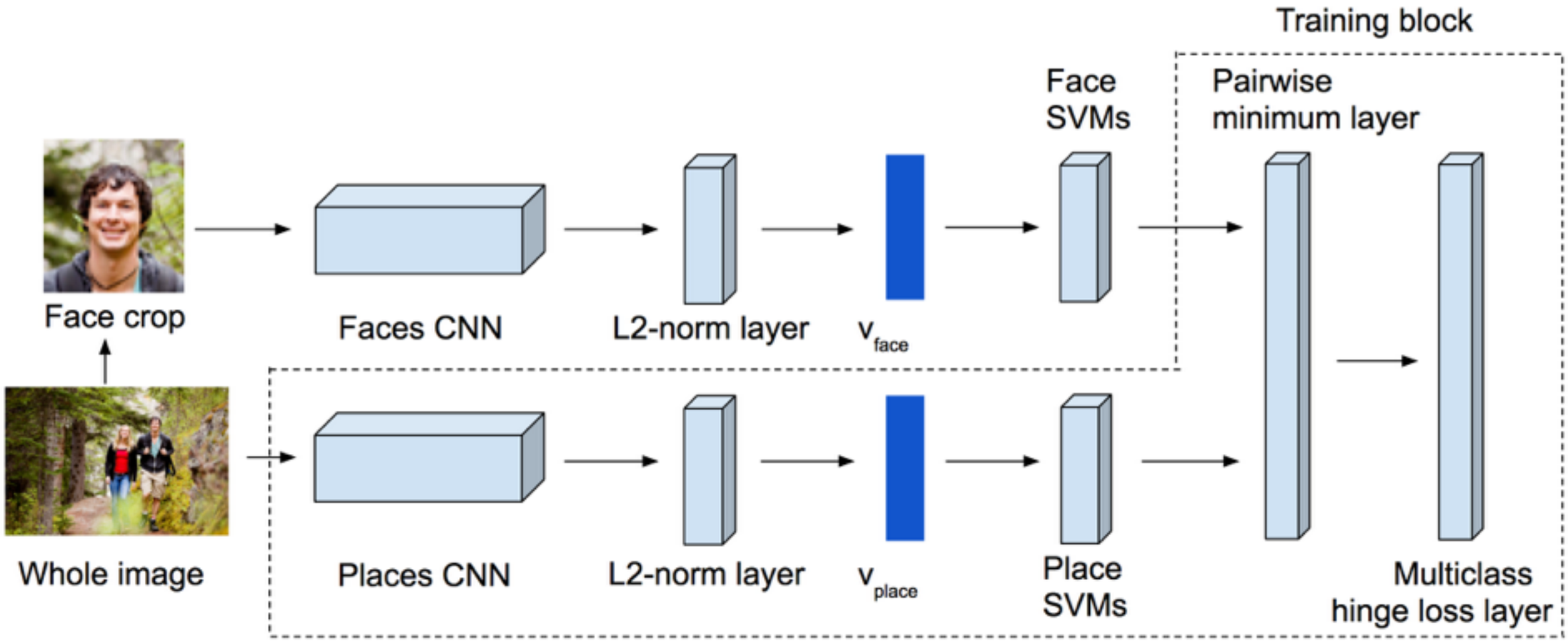
Barack Obama on the **beach**

Arian Foster in the **stadium**

Abbie Cornish at the **ice skating rink**

Hybrid CNN

A stream for face identity and a stream for scene type



Where do we get the training data?

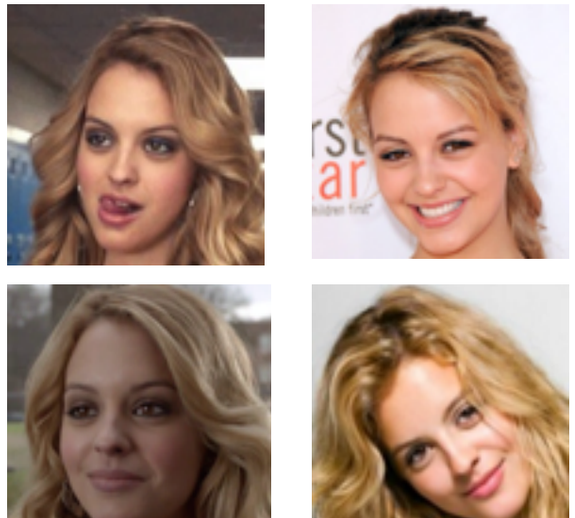
WarpNet: Weakly Supervised Matching for Single-view Reconstruction

Scene dataset + people dataset = people in scenes

MIT Places

“airport terminal”

VGG Faces



“Gage Golightly”

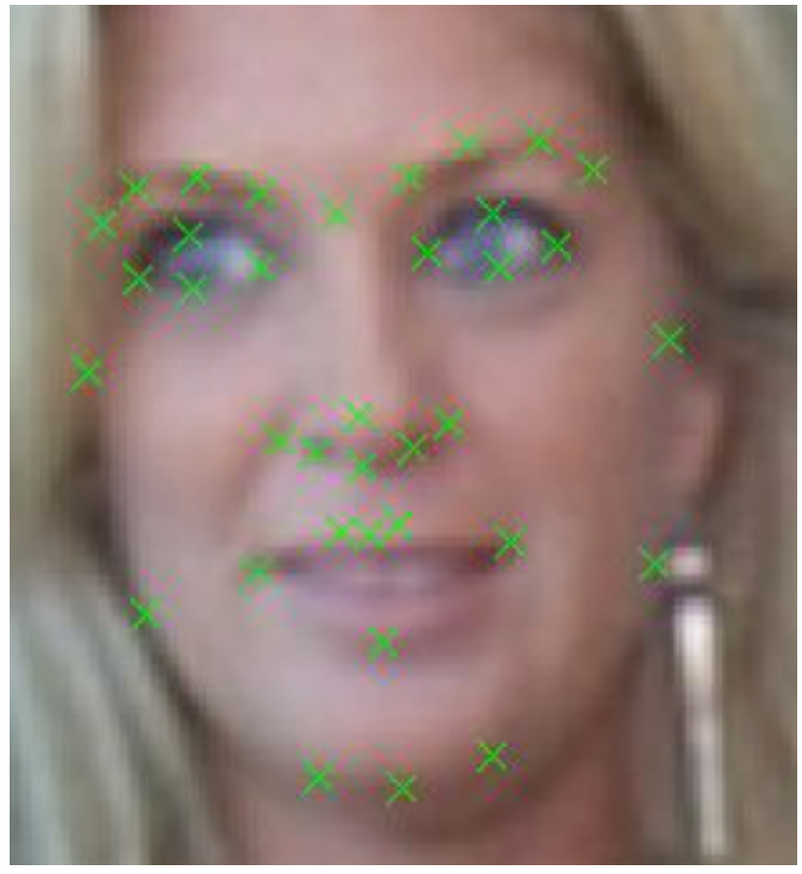
Synthetic

“Gage Golightly at the airport terminal”

[Y. Zhong, R. Arandjelović, A. Zisserman 2016]

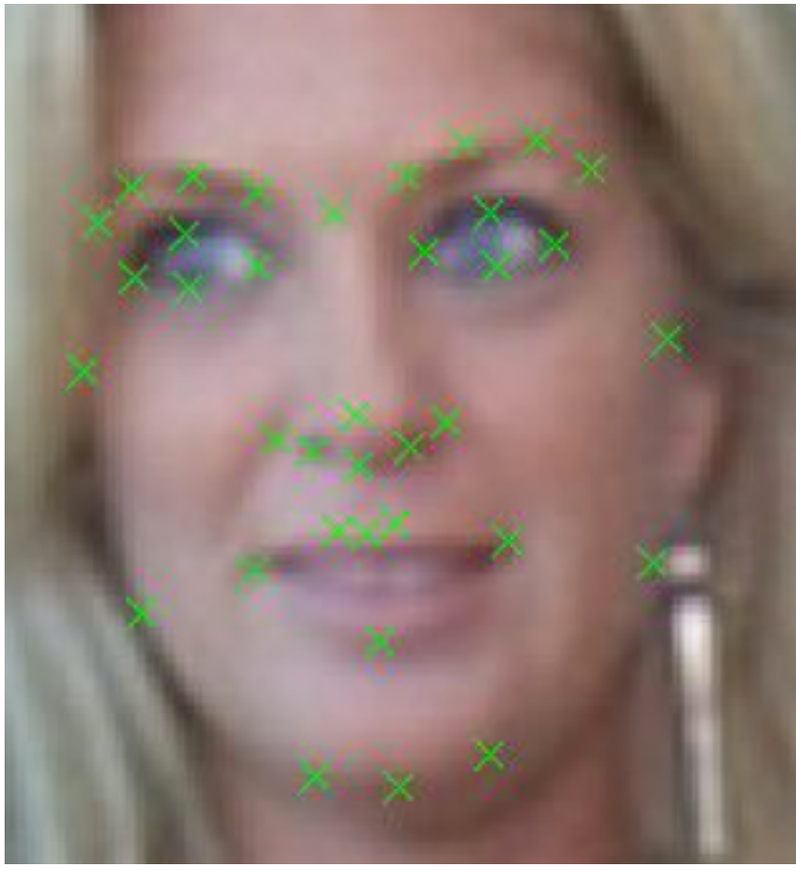
Step 1

Face detection & keypoint estimation



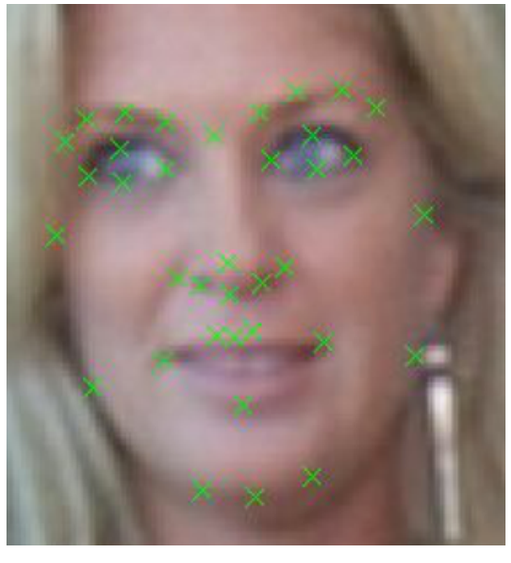
Step 2

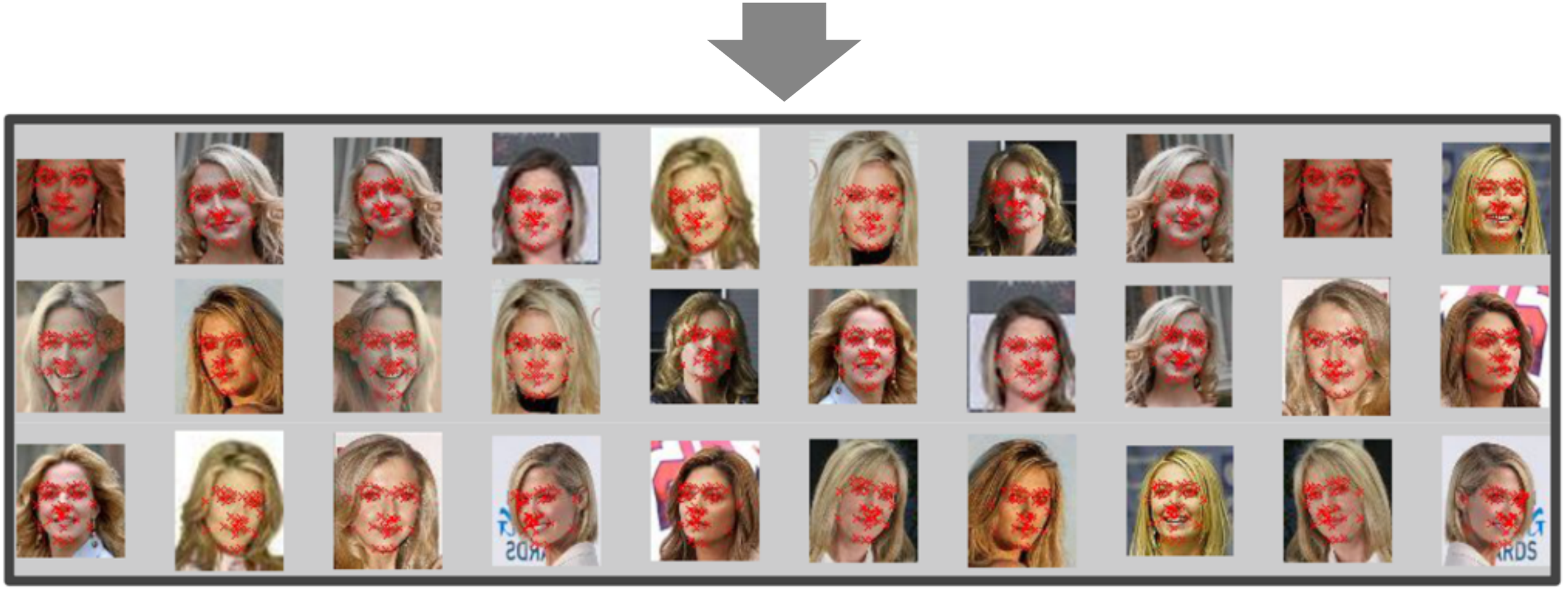
Face search by feature similarity



Step 3

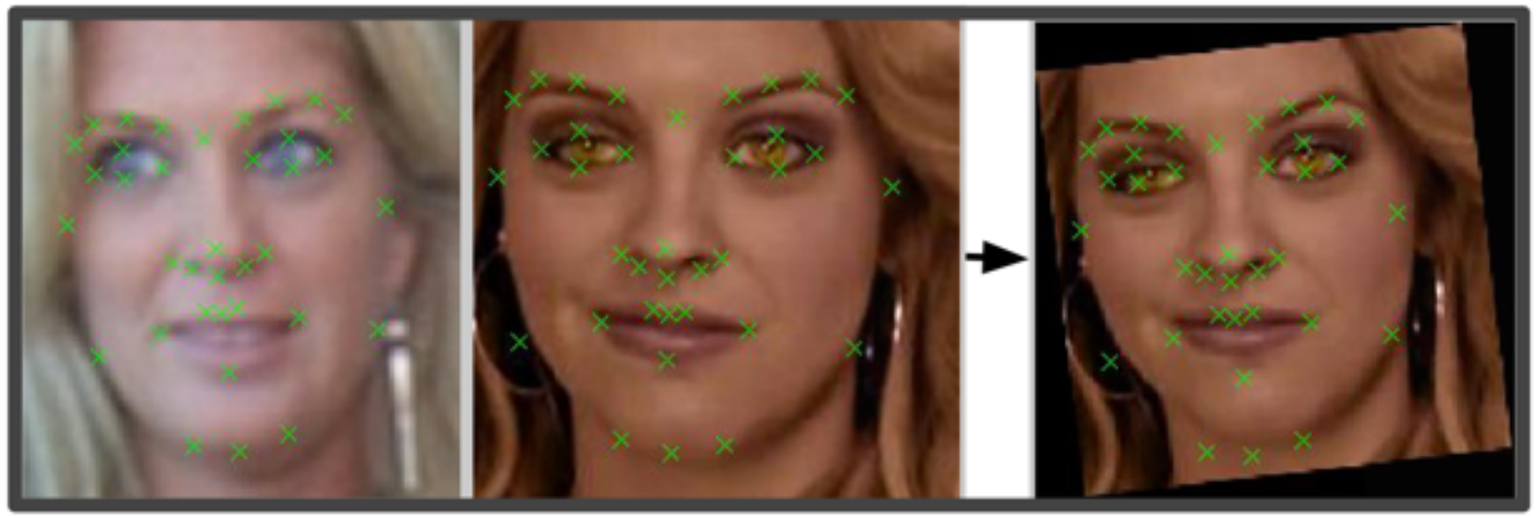
Rerank by pose similarity



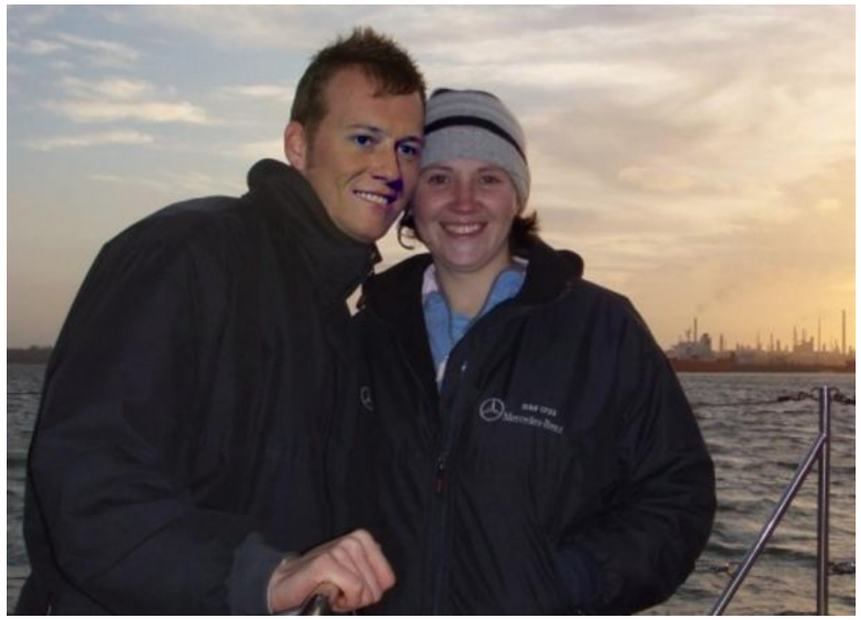


Steps 4,5,6

Alignment, Poisson editing, CNN verification

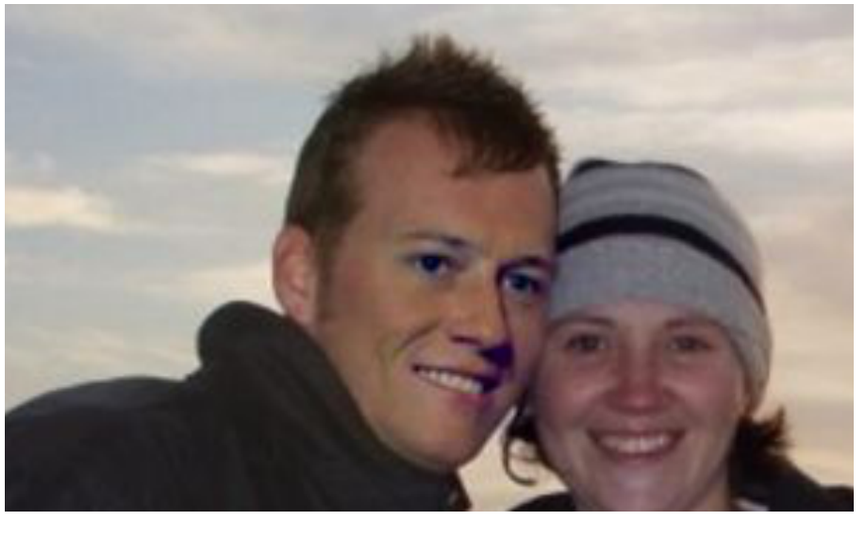


Example substitutions








Example substitutions

Celebrity in Places dataset

Training: synthetic "celebrity in places" + distractors

- No. images 16k (+ 58k distractors)
- No. celebrities 500
- No. places 16

Evaluation: real "celebrity in places" + distractors

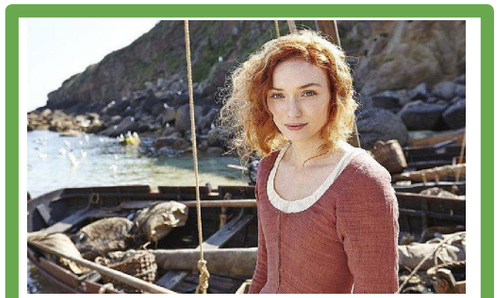
- No. images 1k (+ 58k distractors)
- No. queries 1015

Descriptors	Faces in places (mAP)		Places only (mAP)
	unseen	seen	
Baseline	0.381	0.325	0.381
Our CNN	0.640	0.577	0.514

Retrieval examples

Audrey Hepburn at the **golf course**

Eleanor Tomlinson on the **boat**



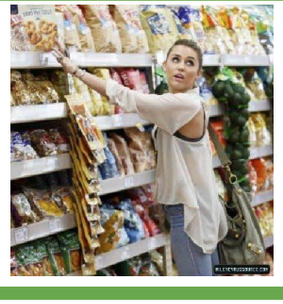
Anthony Rapp on **stage**

Barack Obama on the **beach**

Arian Foster in the **stadium**

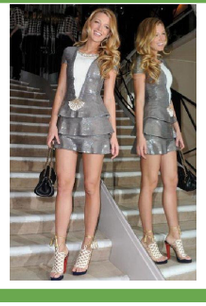
Abbie Cornish at the **ice skating rink**

Miley Cyrus in the **supermarket**



Kuno Becker in the **hospital**

Blake Lively at the **staircase**



Chris Martin at the **airport terminal**

Summary

Modern CNNs are powerful but data hungry

- Annotated training data is often a deal breaker
- Learning with less supervision is paramount
- In the mean time, synthetically-augmented data can help

Synth Text

- A purely synthetic dataset of images of words
- State-of-the-art text recognition in real scenes

Synth Scene Text

- Synthetic text embedded in real scenes
- Automatic blending of synthetic and real elements
- Use deep learning to understand the 3D scene geometry
- State-of-the-art text detection in real scenes

Synthetically-augmented data for deep text spotting

In this talk I will discuss synthetic data augmentation as a strategy for generating large quantities of supervised training data for deep learning. This approach combines two common methods: data augmentation, which generates new training images by transforming existing ones, and synthetic data generation, which creates training images using computer graphics. Synthetic data augmentation transforms real images by inserting virtual objects obtained using computer graphics.

I will discuss the importance of realism in synthetic data augmentation, and show how computer vision techniques such as monocular depth estimation can be used to automatically insert virtual objects in a way which is geometrically consistent with a given scene geometry. I will show that by using such techniques it is possible to construct datasets that are orders of magnitude larger than manually collected ones while being sufficiently realistic for the purpose of machine learning for image understanding.

I will demonstrate these ideas in the context of text spotting. First, I will introduce a synthetic dataset, Synth Text, and show how this can be used to train deep state-of-the-art neural network for text recognition in natural scenes without using any real image. Then, I will introduce a synthetically-augmented dataset, Synth Scene Text, and use the latter to train deep networks for text detection in natural scenes.