

Deep Learning for Vision

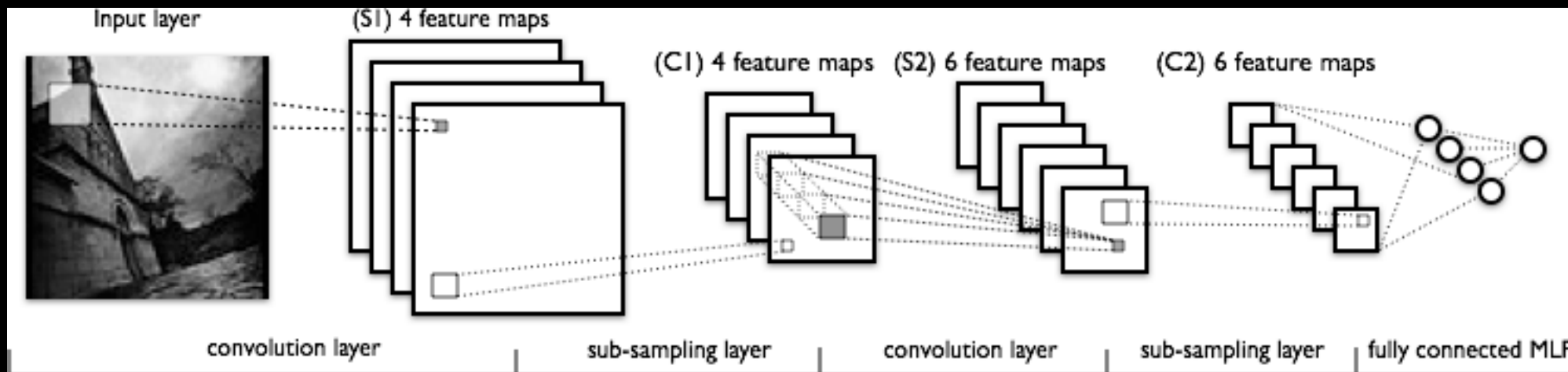
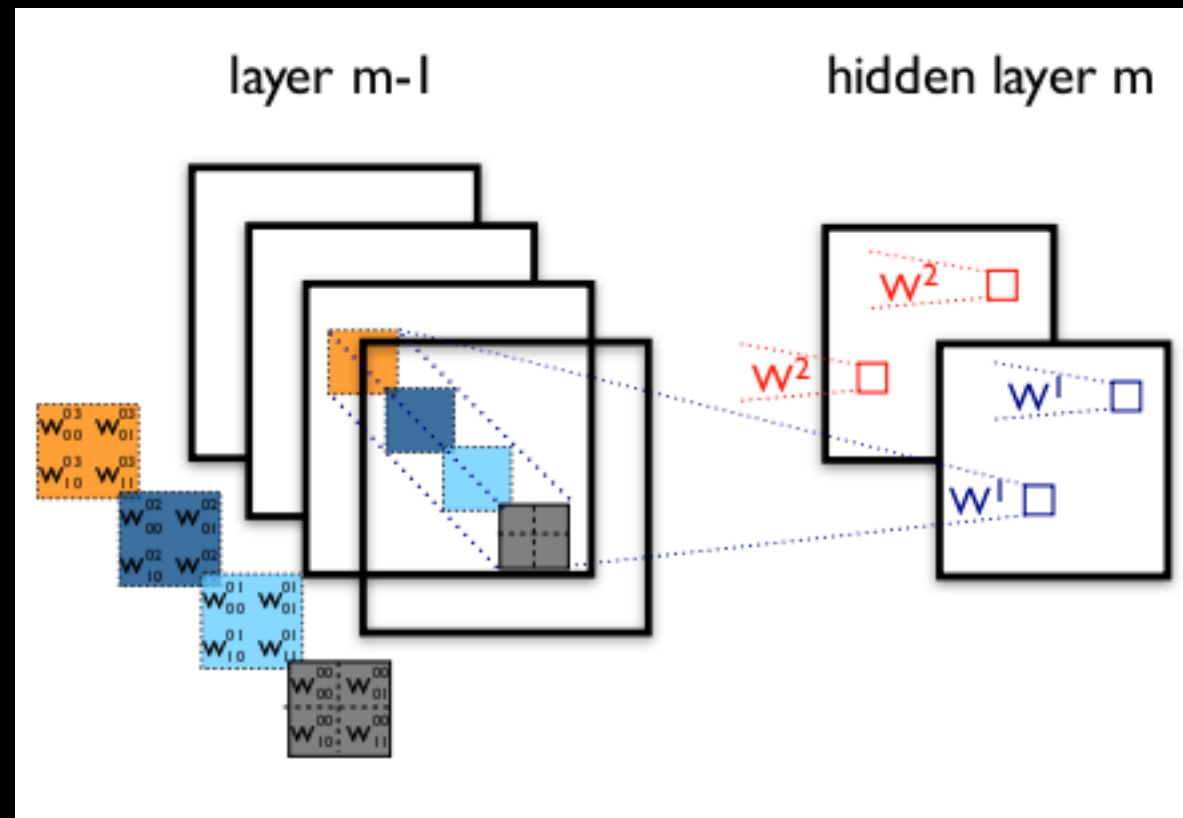
Presented by Kevin Matzen

Quick Intro - DNN

- Feed-forward
- Sparse connectivity (layer to layer)
- Different layer types
- Recently popularized for vision
[Krizhevsky, et. al. NIPS 2012]

The Layers

- Convolution
- Fully connected
- Pooling
- Neuron activation function
- Normalization
- Loss functions
- Image processing



deeplearning.net/tutorial/lenet.html

Software

- code.google.com/p/cuda-convnet/
[nvidia gpu]
- github.com/UCB-ICSI-Vision-Group/decaf-release/
[deprecated; cpu-only]
- caffe.berkeleyvision.org
[cpu; nvidia gpu]
- [research.google.com/archive/
large_deep_networks_nips2012.html](https://research.google.com/archive/large_deep_networks_nips2012.html)
[proprietary; distributed system]

DeepPose: Human Pose Estimation via Deep Neural Networks

Alexander Toshev, Christian Szegedy – CVPR 2014

DeepFace: Closing the Gap to Human-Level Performance in Face Verification

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, Lior Wolf – CVPR 2014

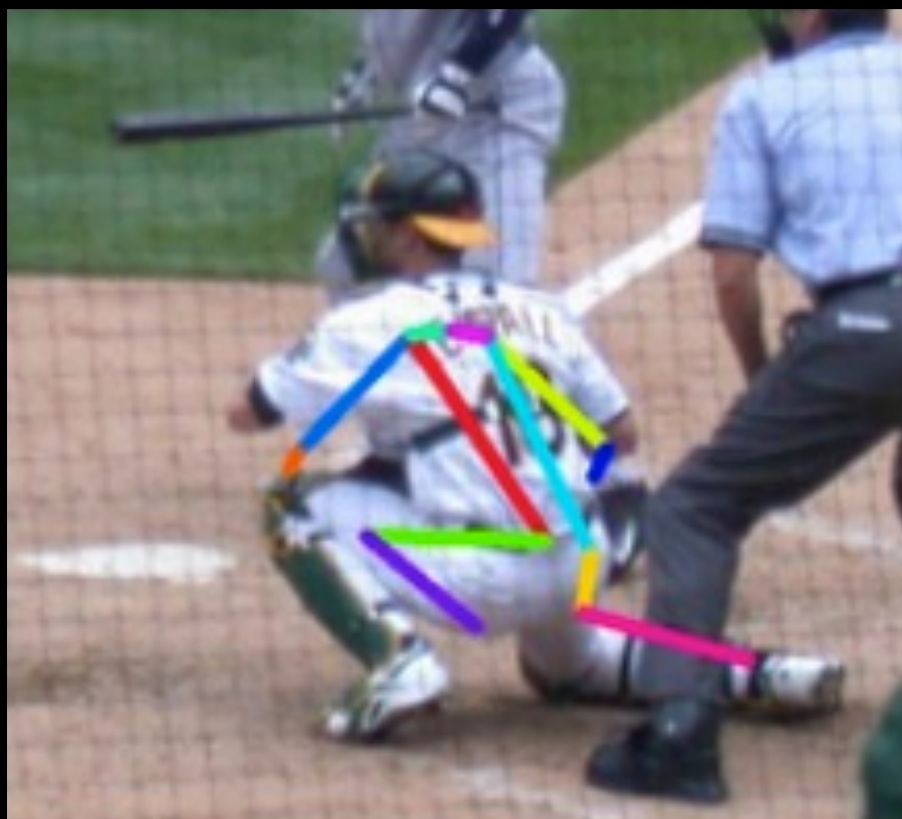
DeepPose: Human Pose Estimation via Deep Neural Networks

Alexander Toshev, Christian Szegedy – CVPR 2014

DeepFace: Closing the Gap to Human-Level Performance in Face Verification

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, Lior Wolf – CVPR 2014

Input: Uncropped photo
Output: Joint locations



Pipeline

1. Person detection
2. Joint position regression
3. Joint refinement

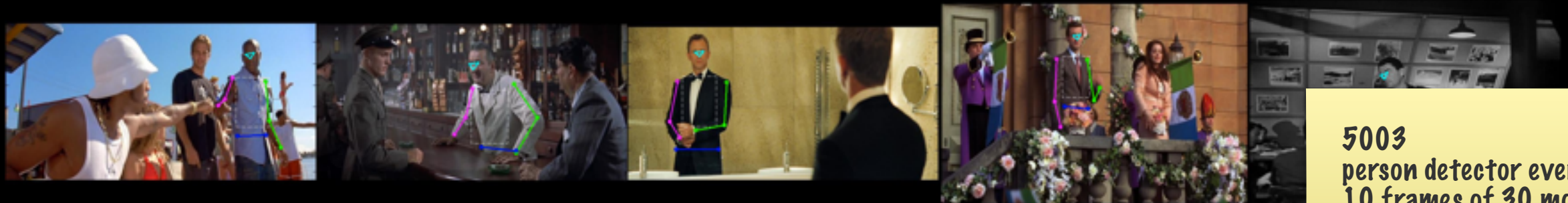
Datasets

Leeds Sports Pose (LSP) [Johnson, et. al. BMVC 2010]



14 joint locations
2000
main person - 150 px

Frames Labeled in Cinema (FLIC) [Sapp, et. al. CVPR 2013]



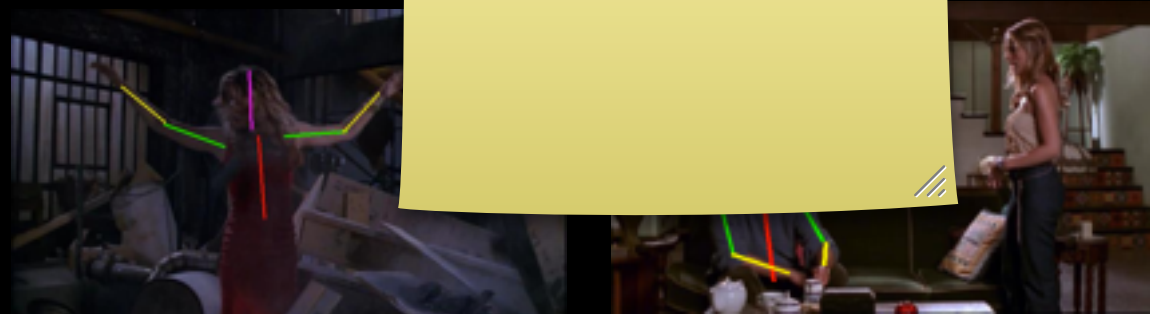
5003
person detector every
10 frames of 30 movies
20k candidates
mturk
10 upperbody joints

Image Parse [Ramanan NIPS 2006]

305 images
similar to leads
includes casual photos

Buffy Stickmen

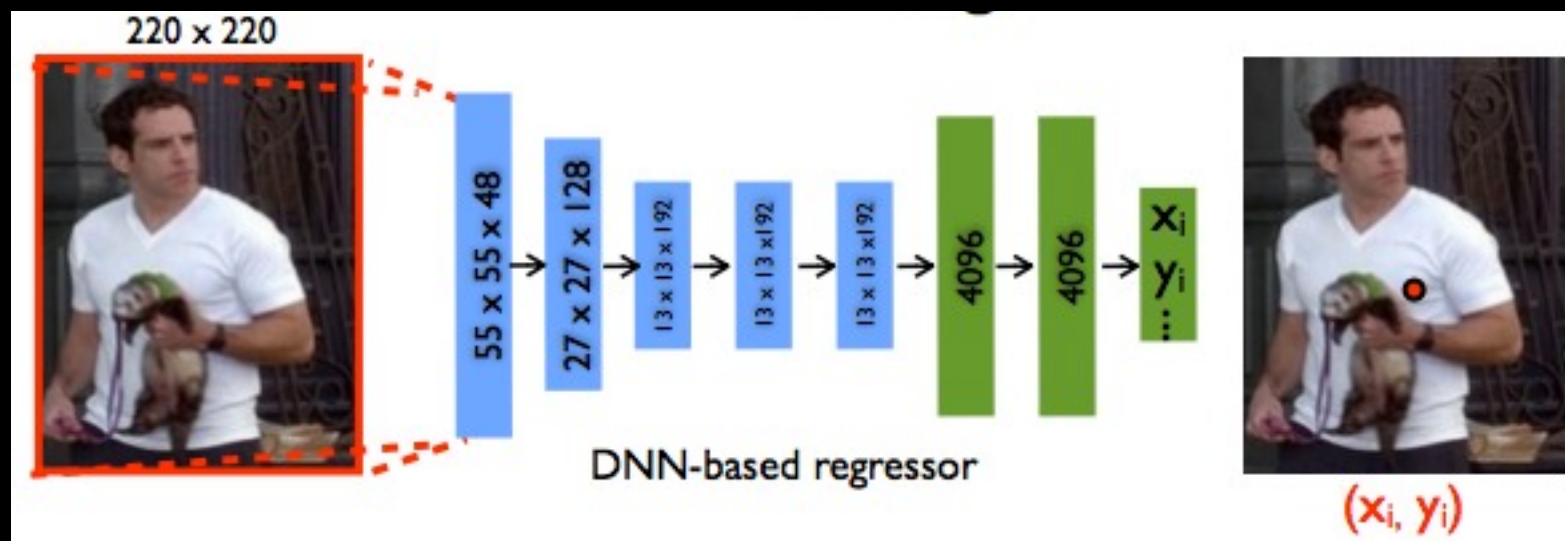
748 frames



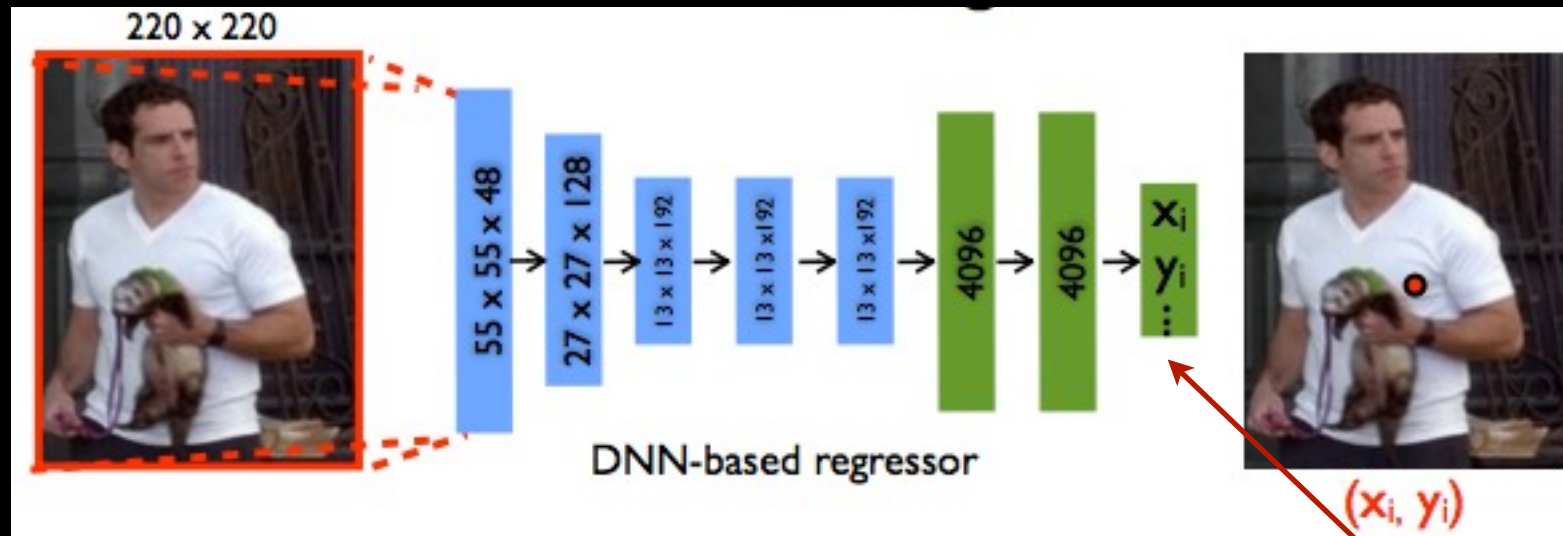
Person Detection

- Input: Uncropped image
- Output: Cropped image

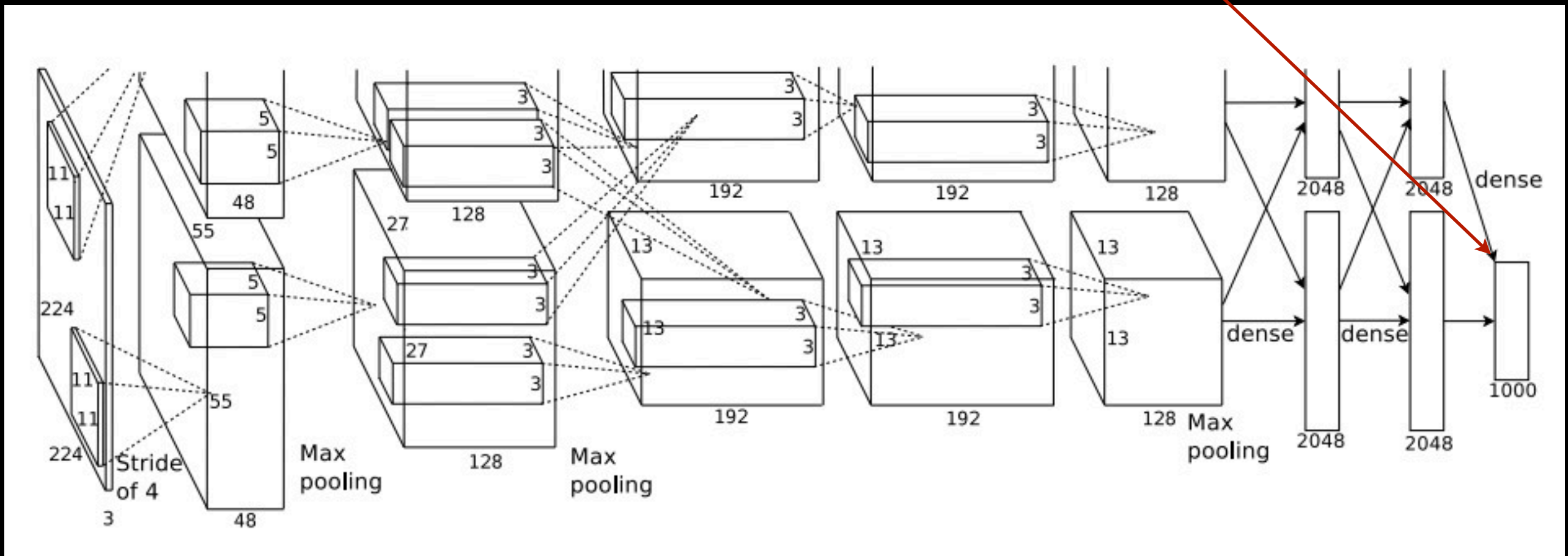
- LSP dataset - No person detector
- FLIC dataset - Enlarged face detector

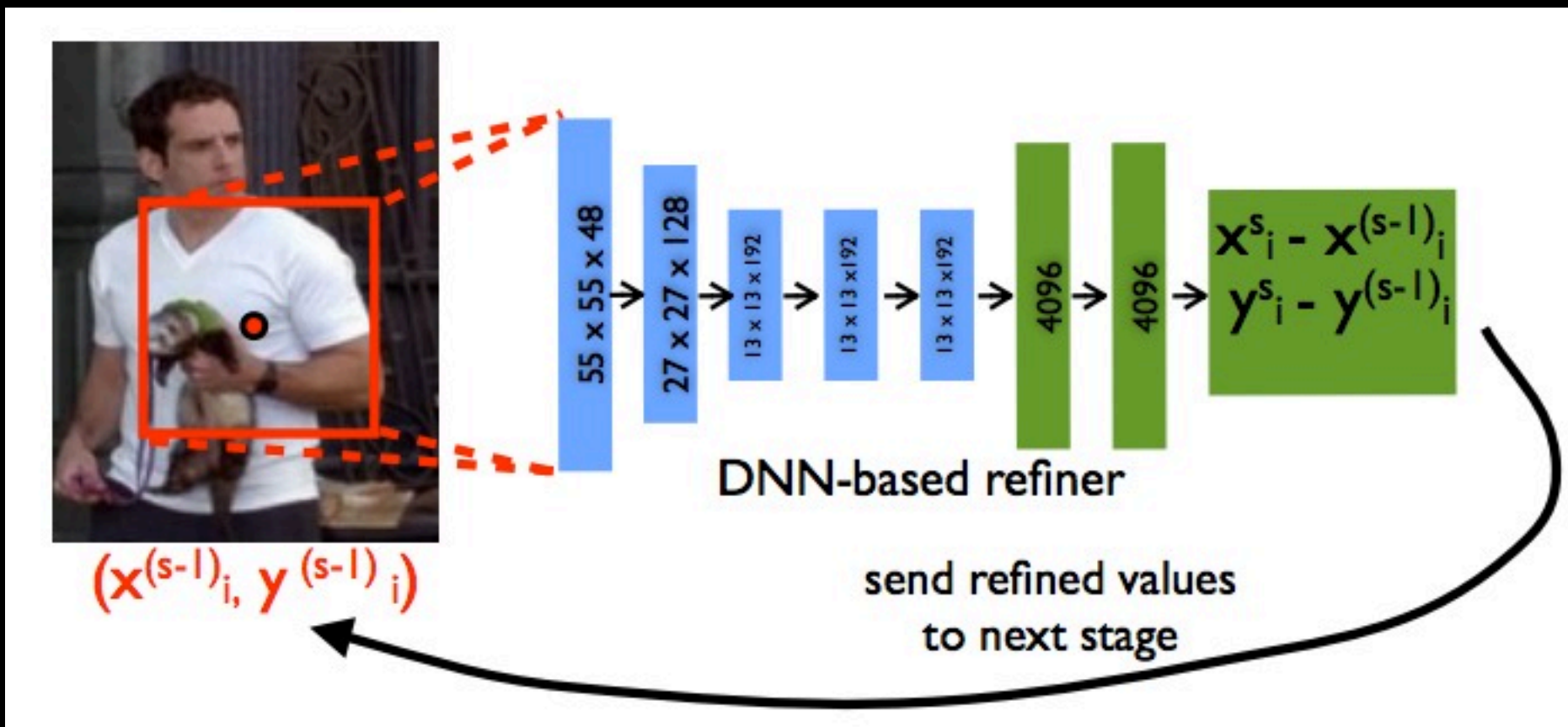


$$\arg \min_{\theta} \sum_{(x,y) \in D_N} \sum_{i=1}^k \|\mathbf{y}_i - \psi_i(x; \theta)\|_2^2$$



Main difference





Runtime

- 0.1s per image - 12 cores (SotA - 1.5s, 4s)
- Training stage 0 - 3 days
- Training refinement - 7 days each

Evaluation

- Percentage of Correct Parts (PCP)
 - Correct if predicted limb is within 1/2 of correct limb length
- Percentage of Detected Joints (PDJ)
 - Predicted and correct joints are within some factor of torso diameter

Method	Arm		Leg		Ave.
	Upper	Lower	Upper	Lower	
DeepPose-st1	0.5	0.27	0.74	0.65	0.54
DeepPose-st2	0.56	0.36	0.78	0.70	0.60
DeepPose-st3	0.56	0.38	0.77	0.71	0.61
Dantone et al. [2]	0.45	0.25	0.65	0.61	0.49
Tian et al. [21]	0.52	0.33	0.70	0.60	0.56
Johnson et al. [11]	0.54	0.38	0.75	0.66	0.58
Wang et al. [22]	0.565	0.37	0.76	0.68	0.59
Pishchulin [15]	0.49	0.32	0.74	0.70	0.56

Table 1. Percentage of Correct Parts (PCP) at 0.5 on LSP for DeepPose as well as five state-of-art approaches.

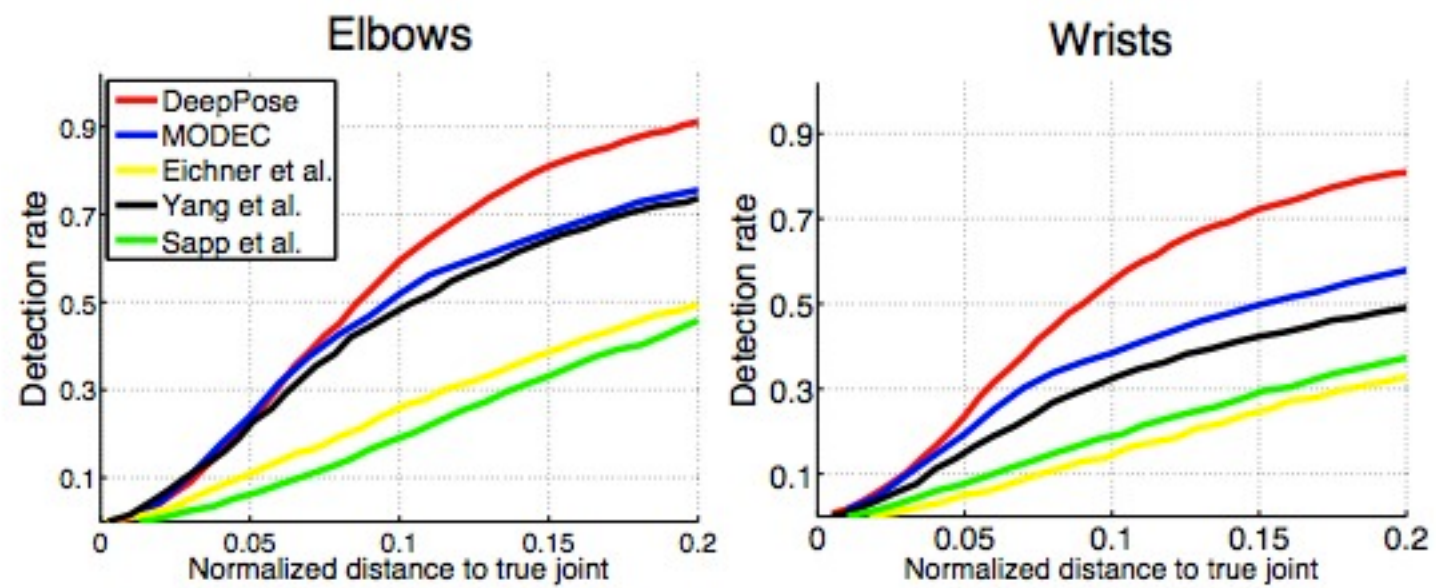


Figure 3. Percentage of detected joints (PDJ) on FLIC for two joints: elbow and wrist. We compare DeepPose, after two cascade stages, with four other approaches.

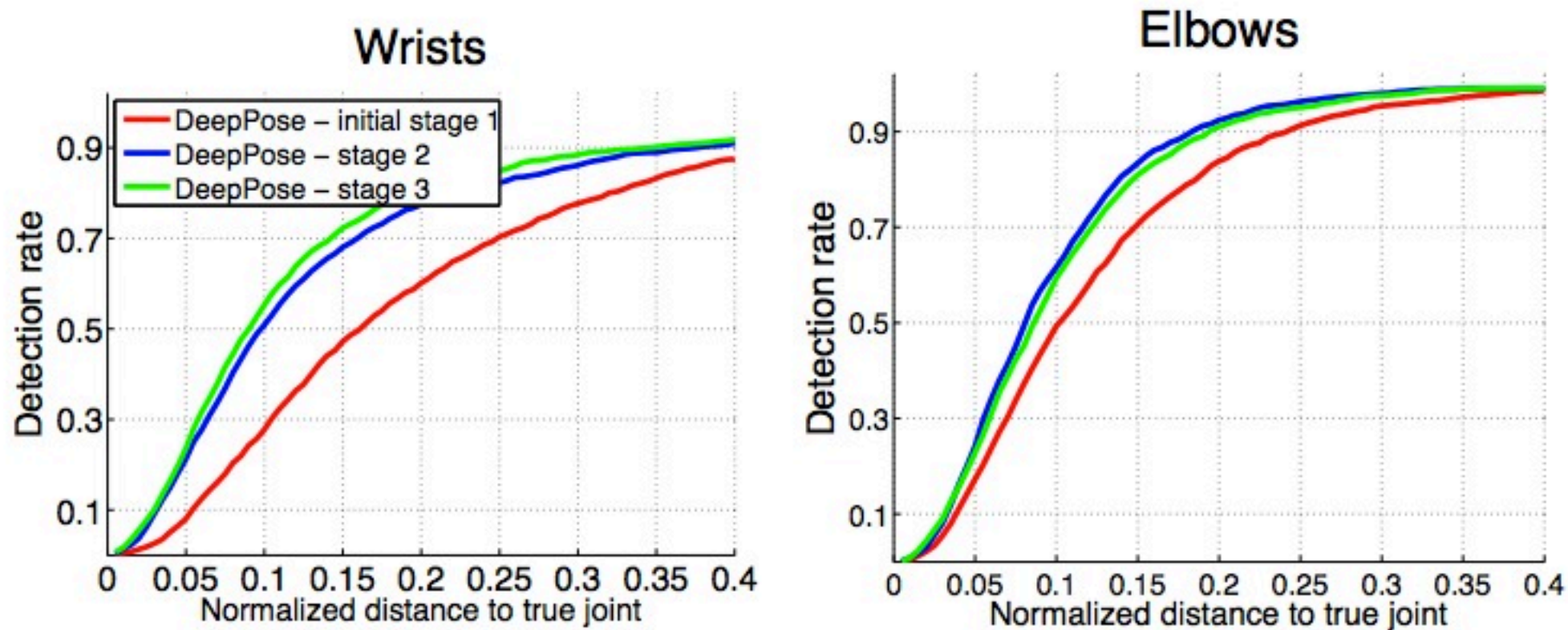


Figure 5. Percent of detected joints (PDJ) on FLIC or the first three stages of the DNN cascade. We present results over larger spectrum of normalized distances between prediction and ground truth.

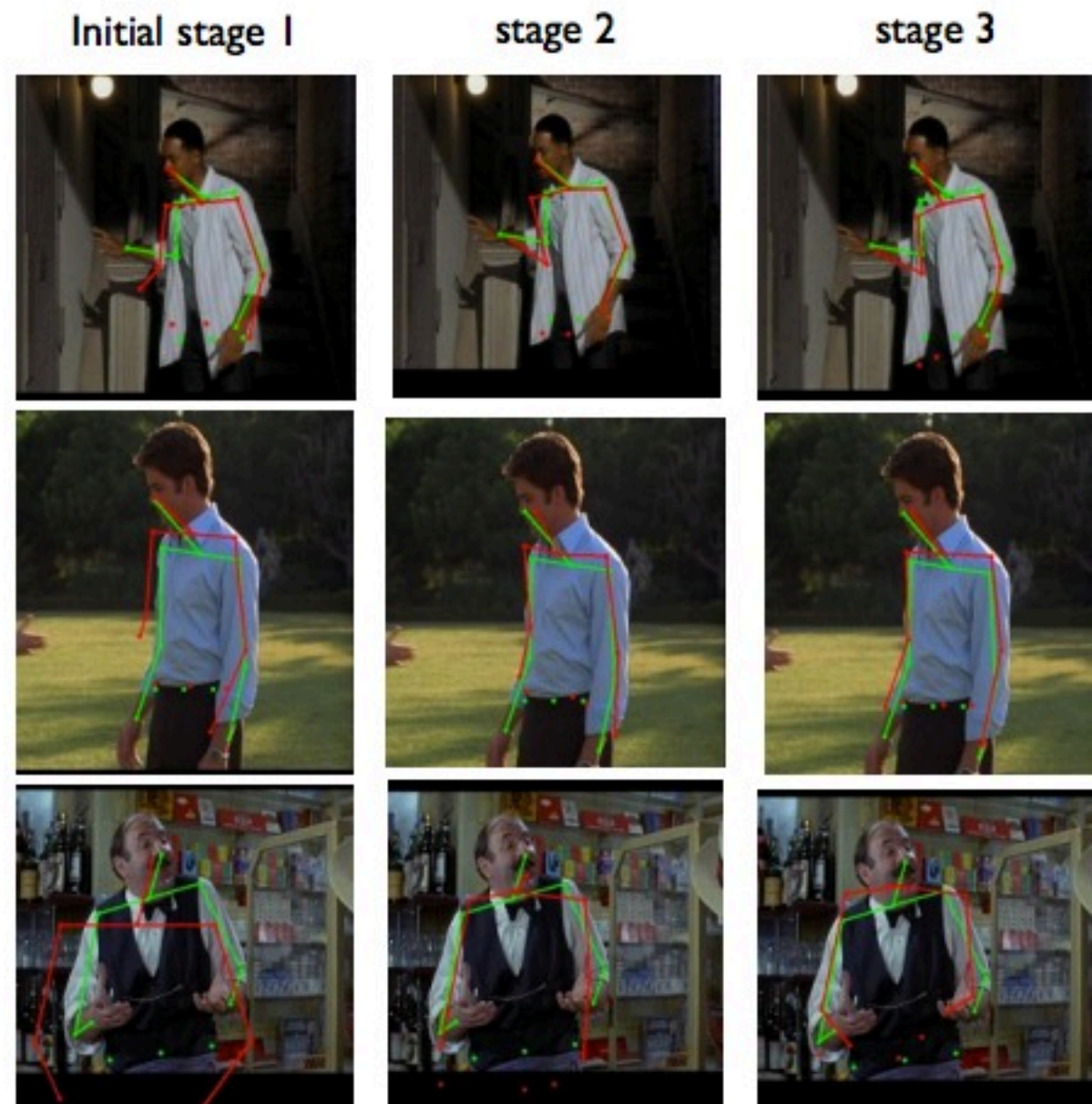


Figure 6. Predicted poses in red and ground truth poses in green for the first three stages of a cascade for three examples.

DeepPose: Human Pose Estimation via Deep Neural Networks

Alexander Toshev, Christian Szegedy – CVPR 2014

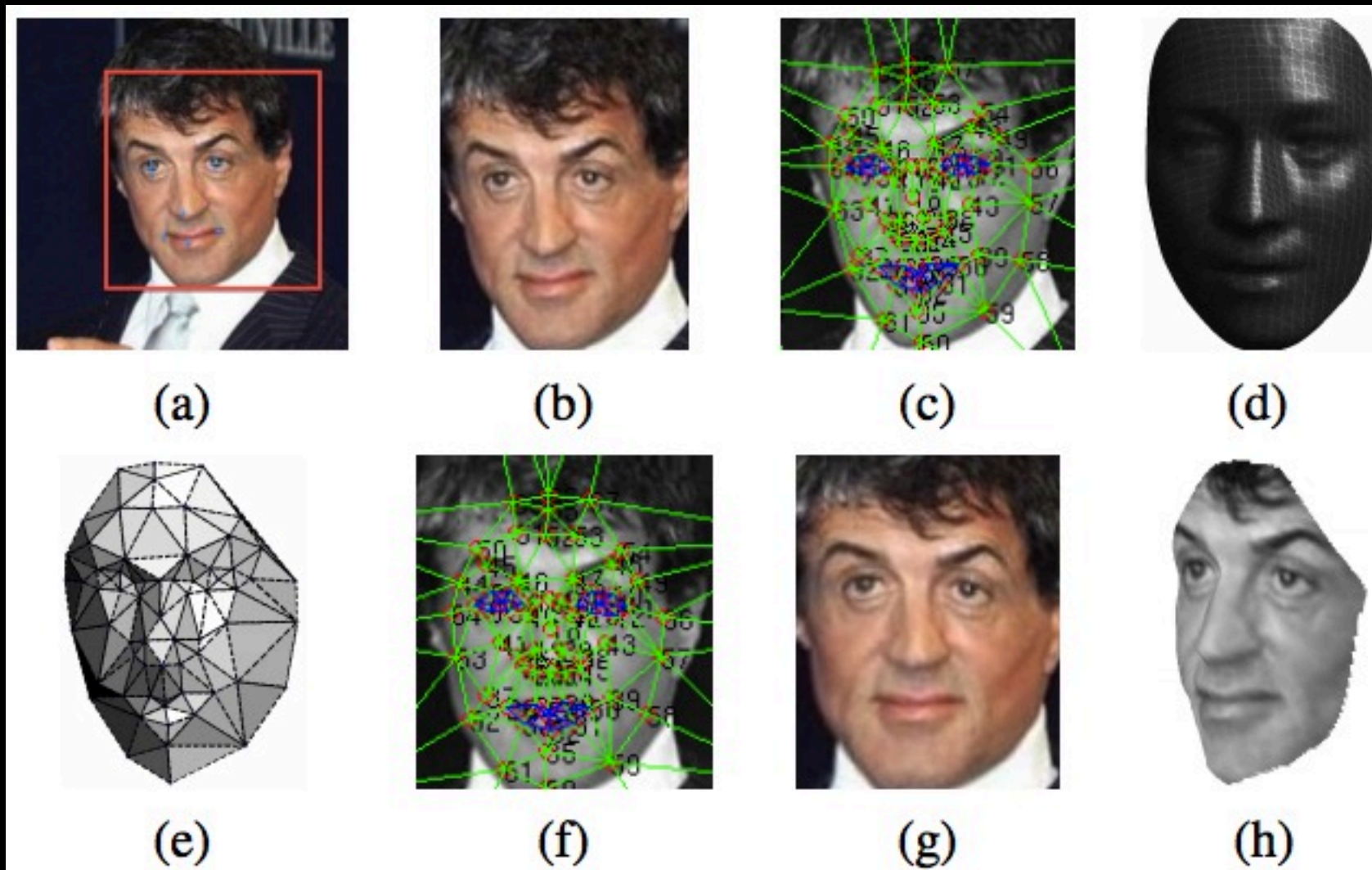
DeepFace: Closing the Gap to Human-Level Performance in Face Verification

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, Lior Wolf – CVPR 2014

Pipeline

- Detect faces
- Correct out-of-plane rotation
- Generate features via CNN
- Classify

Alignment



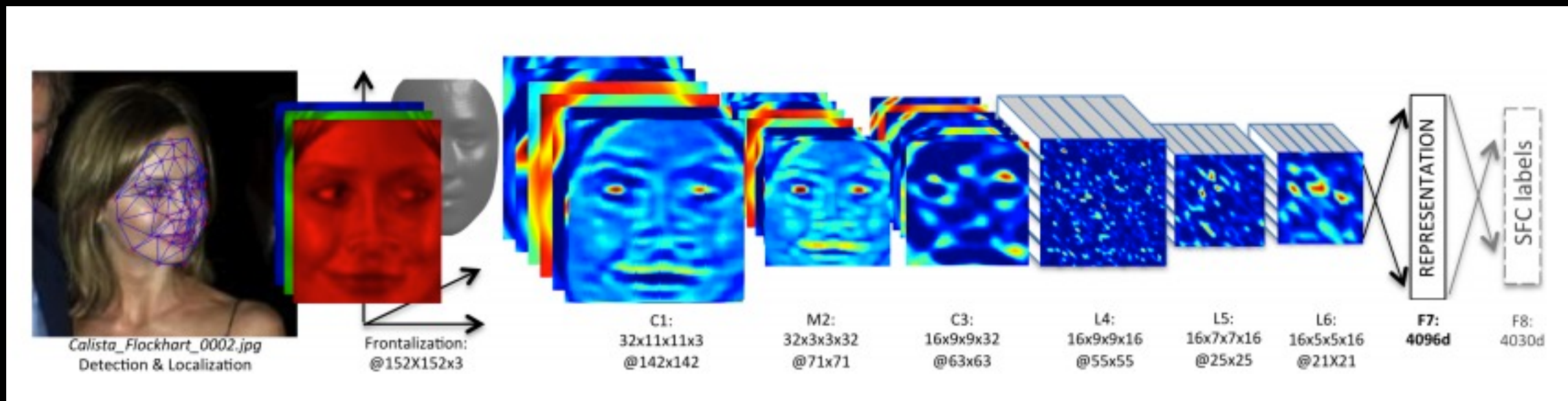
Fiducial Detection

- LBP histograms
- Support Vector Regressor
- Iteratively transform and predict
- 6 fiducial points for 2D alignment
- 67 fiducial points for 3D alignment

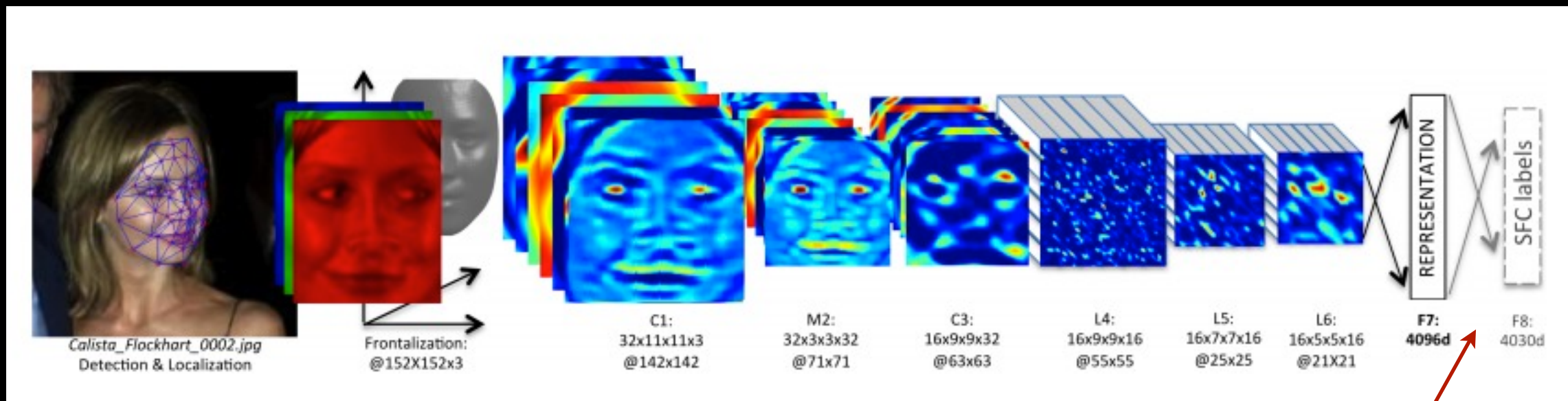
3D Alignment

- Iterative affine camera PnP
- 3D reference - Average mesh of USF Human-ID dataset
- Considers fiducial covariance
- Residuals applied to reference mesh
- Affine warp texture

CNN Architecture

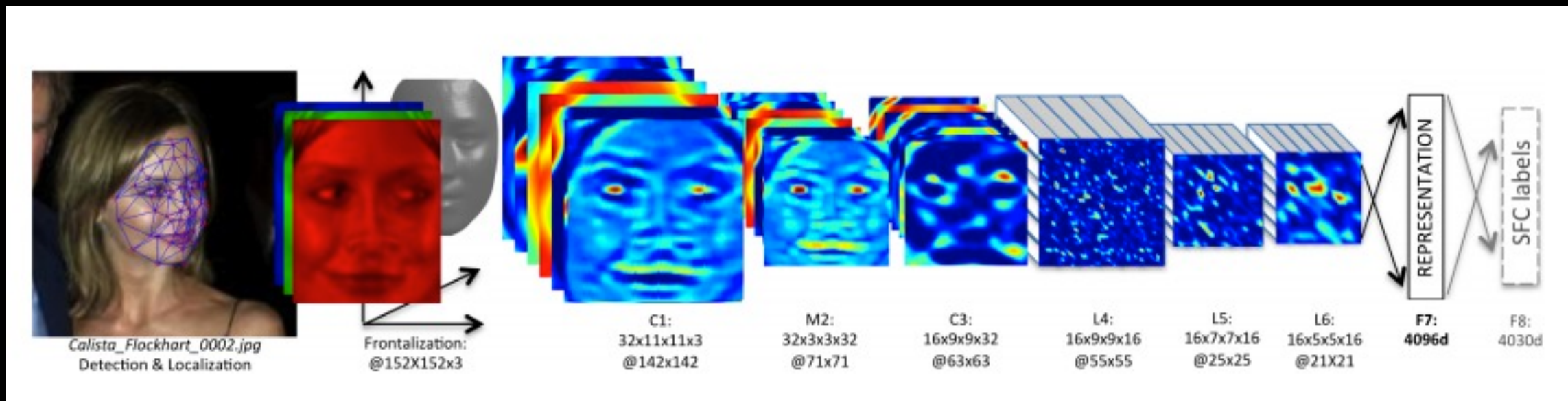


CNN Architecture



Features

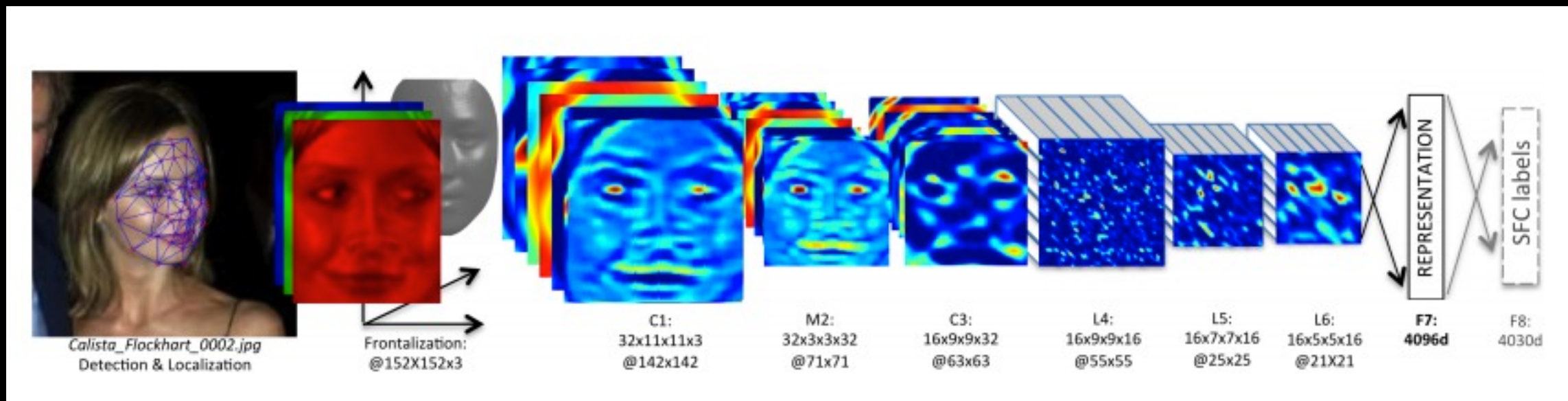
CNN Architecture



weight sharing

no weight sharing

Training



softmax → cross-entropy loss $-\log p_k$

Sparsity

- ReLU nonlinearly - rectified linear unit
 $\max(0, x)$
- 75% model parameters = 0
- Dropout - first fully connected layer

Normalization

- ReLU - unbounded
- Normalize features to $[0, 1]$ based on holdout

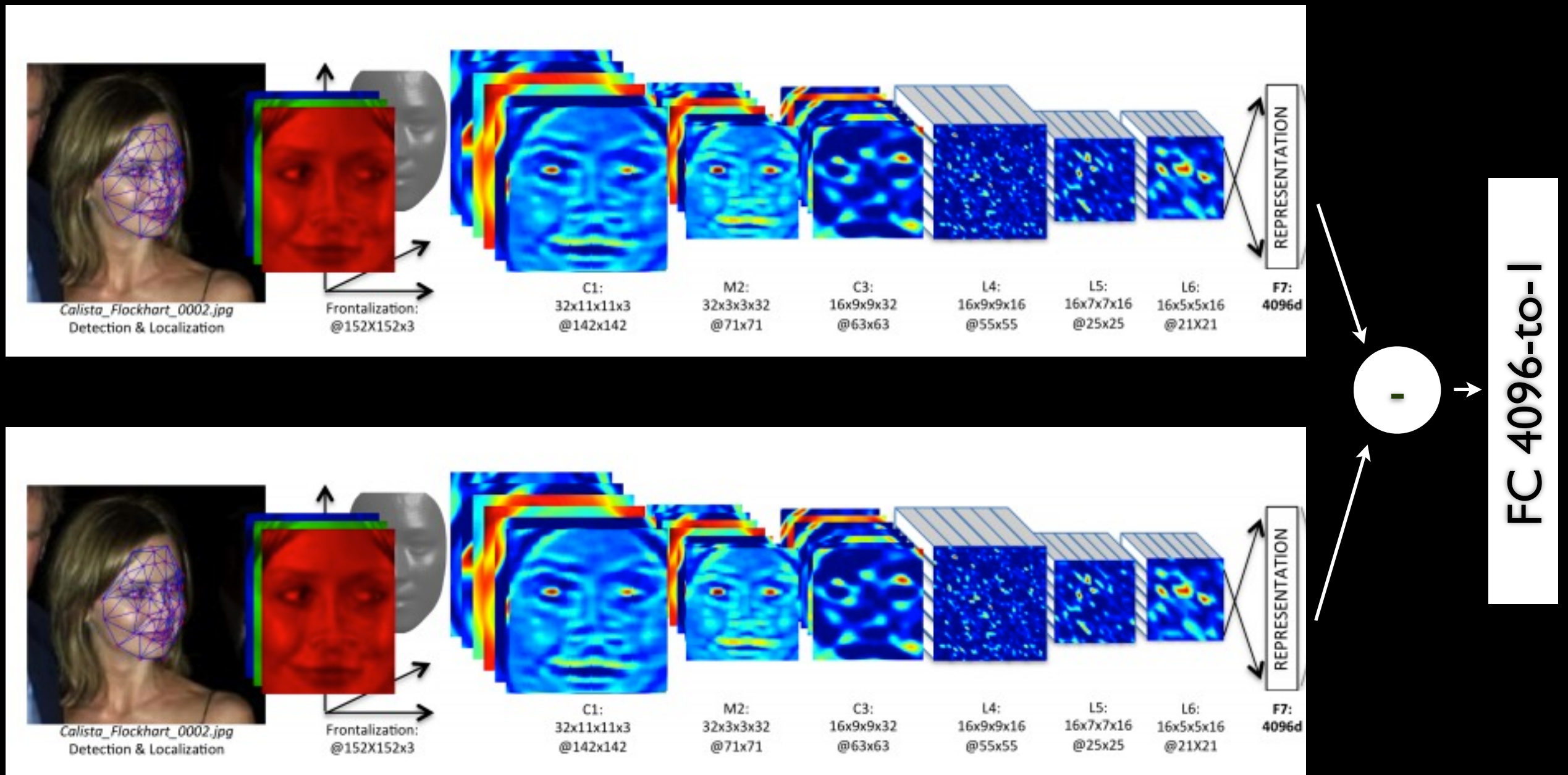
Verification Metrics

- Unsupervised - dot product
- χ^2 similarity
- Siamese network

χ^2 Similarity

- $\chi^2(f_1, f_2) = \sum_i w_i (f_1[i] - f_2[i])^2 / (f_1[i] + f_2[i])$
- weights learned via svm

Siamese Network



Datasets

- Social Face Classification (SFC)
 - Presumably Facebook photos
 - 4.4 mil faces; 4,030 people
 - No overlap with other datasets

Datasets

- Labeled Faces in the Wild (LFW)
 - 13,323 faces; 5,749 celebs
 - 6,000 pairs
 - Restricted protocol - same/not same labels at training
 - Unrestricted protocol - identities during training
 - Unsupervised - no training on LFW

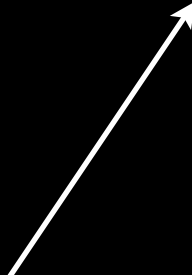
Datasets

- YouTube Faces (YTF)
 - 3,425 videos of 1,595 subjects
 - Subset of celebs from LFW


SFC Training Perf

Network	Error	Network	Error	Network	Error
<i>DF-1.5M</i>	7.00%	<i>DF-10%</i>	20.7%	<i>DF-sub1</i>	11.2%
<i>DF-3.3M</i>	7.22%	<i>DF-20%</i>	15.1%	<i>DF-sub2</i>	12.6%
<i>DF-4.4M</i>	8.74%	<i>DF-50%</i>	10.9%	<i>DF-sub3</i>	13.5%

Reduce data by
omitting people



Reduce data by
omitting examples



Remove layers
from network



LFW Perf

Method	Accuracy	Protocol
Joint Bayesian [6]	0.9242 \pm 0.0108	restricted
Tom-vs-Pete [4]	0.9330 \pm 0.0128	restricted
High-dim LBP [7]	0.9517 \pm 0.0113	restricted
TL Joint Bayesian [5]	0.9633 \pm 0.0108	restricted
DeepFace-single	0.9592 \pm 0.0092	unsupervised
DeepFace-single	0.9700 \pm 0.0087	restricted
DeepFace-ensemble	0.9715 \pm 0.0084	restricted
DeepFace-ensemble	0.9725 \pm 0.0081	unrestricted
Human, cropped	0.9753	

Network	Error (<i>SFC</i>)	Accuracy (<i>LFW</i>)
<i>DeepFace-gradient</i>	8.9%	0.9582 \pm 0.0118
<i>DeepFace-align2D</i>	9.5%	0.9430 \pm 0.0136
<i>DeepFace-Siamese</i>	NA	0.9617 \pm 0.0120

Runtime

- 0.18 s - feature extraction (1 core; 2.2 GHz)
- 0.05 s - alignment
- 0.33 s - total

Questions?