CSE 252D: Advanced Computer Vision

Manmohan Chandraker

Lecture 11: Human Pose Estimation

Virtual classrooms

- Virtual lectures on Zoom
  - Only host shares the screen
  - Keep video off and microphone muted
  - But please do speak up (remember to unmute!)
  - Slides uploaded on webpage just before class
- Virtual interactions on Zoom
  - Ask and answer plenty of questions
  - “Raise hand” feature on Zoom when you wish to speak
  - Post questions on chat window
  - Happy to try other suggestions!
- Lectures recorded and upload on Canvas
  - Available under “My Media” on Canvas

Overall goals for the course

- Introduce the state-of-the-art in computer vision
- Study principles that make them possible
- Get understanding of tools that drive computer vision
- Enable one or all of several such outcomes
  - Pursue higher studies in computer vision
  - Join industry to do cutting-edge work in computer vision
  - Gain appreciation of modern computer vision technologies
- This is a great time to study computer vision!

Papers for Fri, May 07

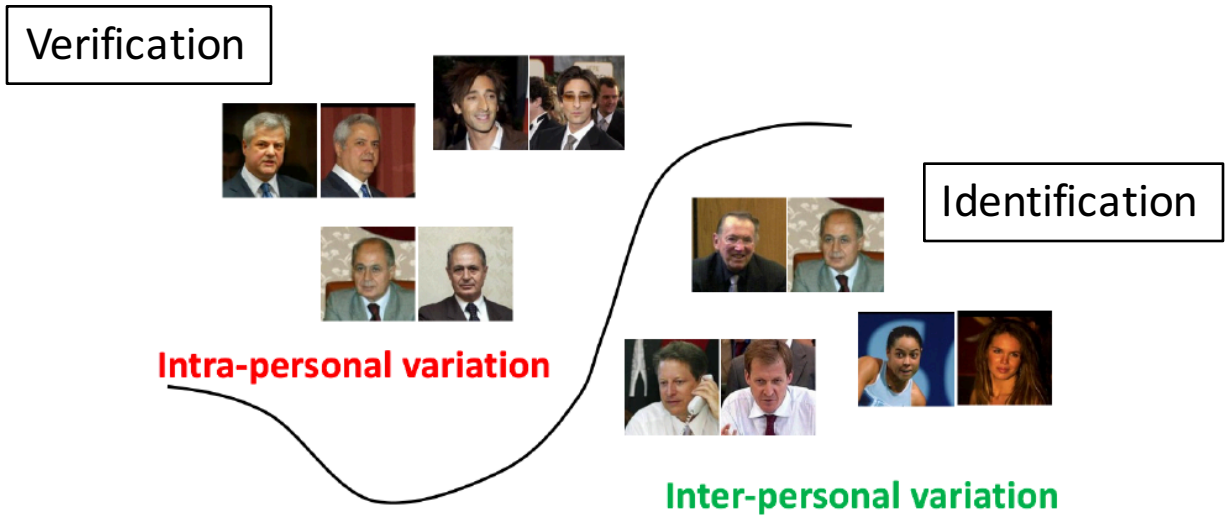
- Deep High-Resolution Representation Learning for Human Pose Estimation
  - https://arxiv.org/abs/1902.09212
- Simple Baselines for Human Pose Estimation and Tracking
  - https://arxiv.org/abs/1804.06208
- OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields
  - https://arxiv.org/abs/1812.08008
- End-to-end Recovery of Human Shape and Pose
  - https://arxiv.org/abs/1712.06584

Papers for Wed, May 12

- ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation
  - https://arxiv.org/abs/1606.02147
- ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation
  - https://ieeexplore.ieee.org/abstract/document/8063438
- Fast-SCNN: Fast Semantic Segmentation Network
  - https://arxiv.org/abs/1902.04502
- Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation
  - https://arxiv.org/abs/1506.04924

Recap

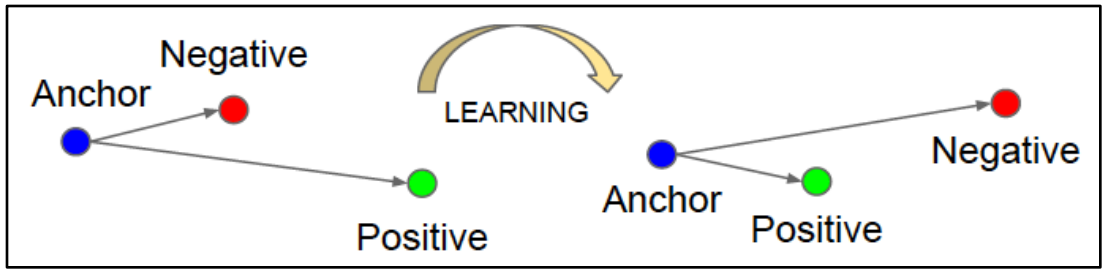
Verification and Identification Signals



- **Identification:**
  - Distinguish images of one identity from another identity
  - Favors large distance between clusters
  - Stronger learning signal, but need not generalize to new identities
- **Verification:**
  - Match two images of an individual across large appearance variations
  - Favors tight clusters for each identity
  - Weaker learning signal, but feature applicable to new identities

Triplet Loss for Training

- Goal for triplet loss:
  - Minimize distance between anchor image x_i^a and a positive x_i^p
  - Maximize distance between anchor x_i^a and a negative x_i^n



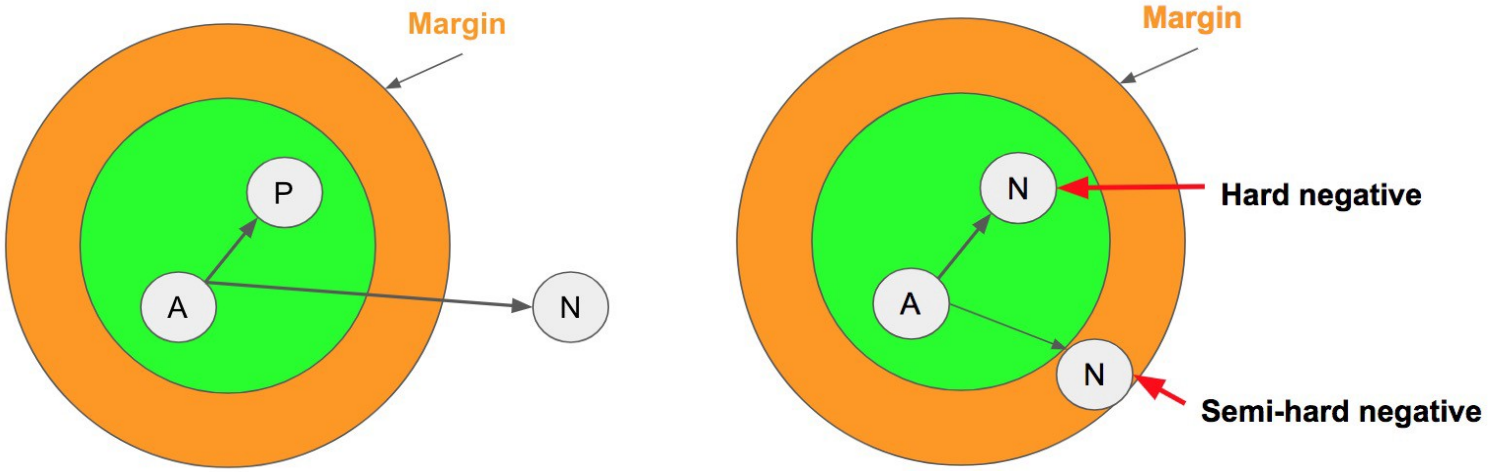
$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2 \quad , \quad \forall \left(f(x_i^a), f(x_i^p), f(x_i^n)\right) \in \mathcal{T}$$

- Total loss to minimize:

$$\sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha\right]_+$$

- Challenge: too many triplets satisfy the margin easily
- Need to select hard examples that are active and improve the model

Triplet Selection



- In practice:
  - Use all anchor-positive pairs, instead of just hard positives
  - Saves expense of searching for hard positive
  - Use **semi-hard negatives** at beginning of training

$$\|f(x_i^a) - f(x_i^p)\|_2^2 < \|f(x_i^a) - f(x_i^n)\|_2^2$$

  - Not hardest negatives, can be within margin, but further than positives
  - Hardest negatives at beginning can cause feature to collapse to $f(x) = 0$

Training Objectives for Face Recognition

- Metric learning

$$\sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

- Softmax :
  - Labeled data (x_i, y_i) for i = 1, ..., m in mini-batch and classes j = 1, ..., n
  - W, b: Weight matrix and bias for last fully-connected layer

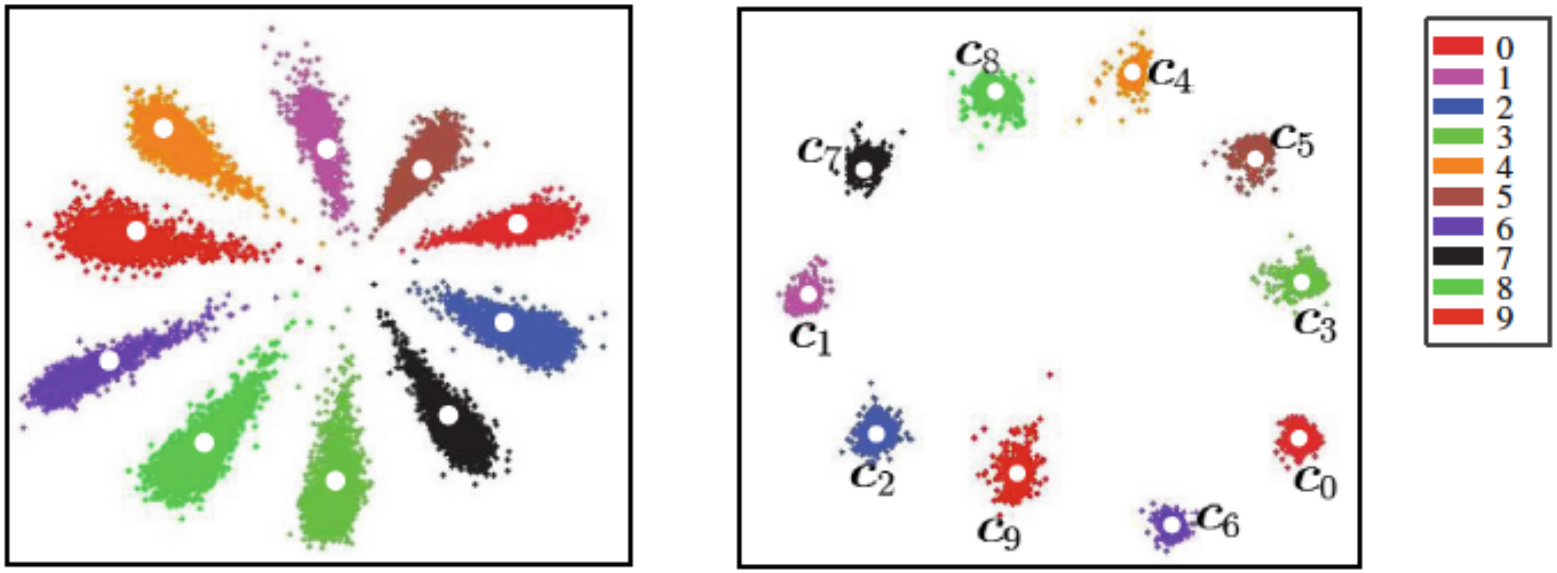
$$\mathcal{L}_S = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \boldsymbol{x}_i + b_j}}$$

Training Objectives: Center Loss

- Softmax pushes features apart for distinct classes
- Introduce a pull term to encourage features of a class to cluster together

$$-\sum_{i=1}^{m}\log\frac{e^{W_{y_i}^T\boldsymbol{x}_i+b_{y_i}}}{\sum_{j=1}^{n}e^{W_j^T\boldsymbol{x}_i+b_j}}+\frac{\lambda}{2}\sum_{i=1}^{m}\|\boldsymbol{x}_i-\boldsymbol{c}_{y_i}\|_2^2$$

  - Enhance discrimination power
  - Sequentially update *W*, CNN parameters and *c* in each mini-batch



[Wen et al., ECCV 2016]

Training Objectives: Angular Softmax

- Demand angular margin between classes, instead of Euclidean margin
  - Consider softmax when W is normalized (and b is 0)

$$\sum_i -\log\left(\frac{e^{\|x_i\|\cos(\theta_{y_i,i})}}{\sum_j e^{\|x_i\|\cos(\theta_{j,i})}}\right)$$

  - Angle between data point x_i and vector W_j : $\theta_{j,i}$

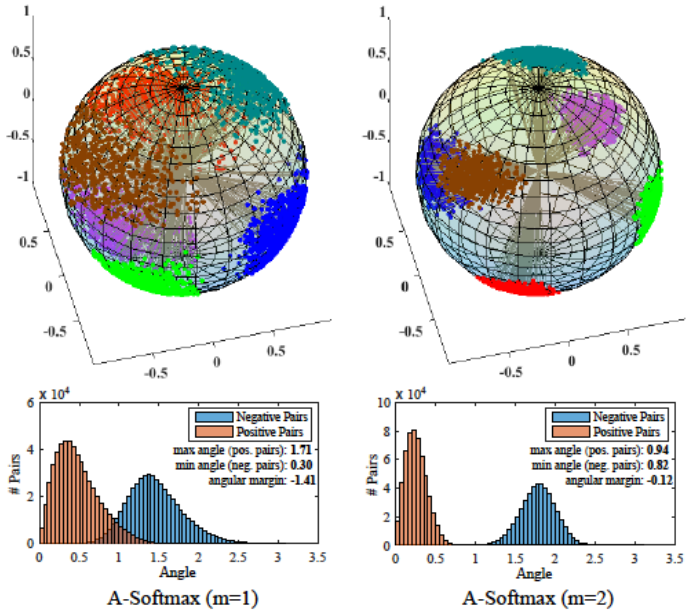
- To correctly classify class 1 against 2:

$$\cos(\theta_1) > \cos(\theta_2)$$

- A-softmax: more stringent lower bound

$$\cos(\theta_1) > \cos(m\theta_1) > \cos(\theta_2)$$

$$\sum_i -\log\left(\frac{e^{\|x_i\|\cos(m\theta_{y_i,i})}}{e^{\|x_i\|\cos(m\theta_{y_i,i})} + \sum_{j\neq y_i} e^{\|x_i\|\cos(\theta_{j,i})}}\right)$$



[Liu et al., SphereFace, CVPR 2017]

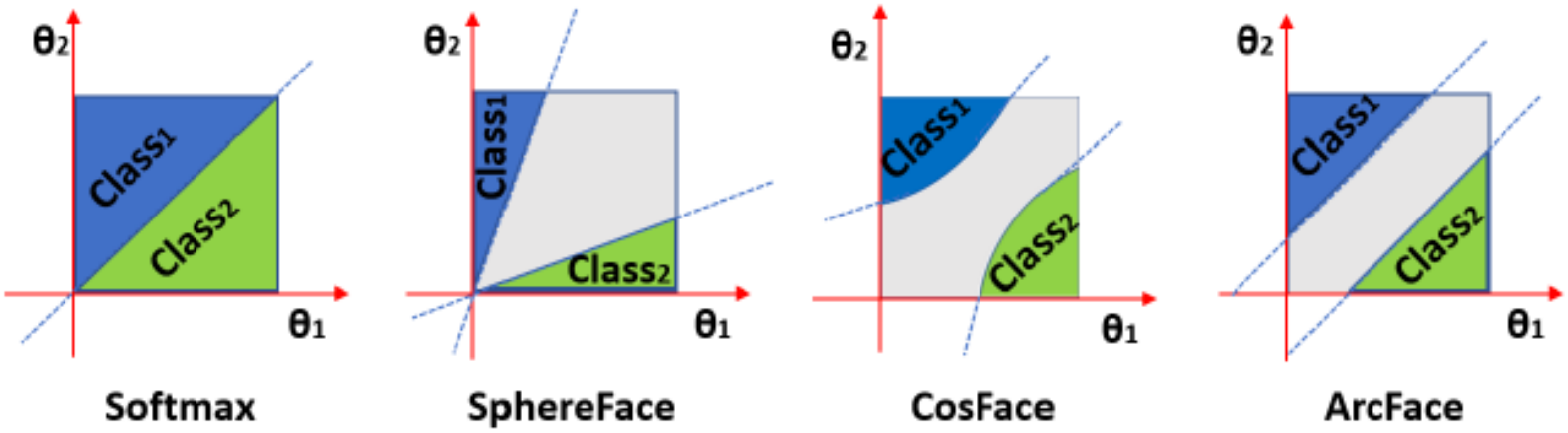
Training Objectives: Additive Angular Margins

- Demand additive angular margin between classes
- CosFace: maintains large angular margin even for visually similar classes

$$\sum_i -\log \frac{e^{s(\cos(\theta_{y_i,i})-m)}}{e^{s(\cos(\theta_{y_i,i})-m)} + \sum_{j \neq y_i} e^{s\cos(\theta_{j,i})}}$$

- ArcFace imposes a margin on $\theta_{j,i}$, more representative of geodesic distance

$$-\sum_{i=1}^{N} \log \frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1, j \neq y_i}^{n} e^{s\cos\theta_j}}$$



[Wang et al., CosFace, CVPR 2018] [Deng et al., ArcFace]

Evaluation of Face Verification

- Given a pair of face images:
  - A squared L2 distance $D(x_i, x_j)$ is used to determine same or different
  - Good embedding: true matches will lie within a small value of $D(x_i, x_j)$

- True accepts: set of face pairs correctly classified as same within a threshold *d*

$$\mathrm{TA}(d) = \{(i,j) \in \mathcal{P}_{\text{same}}, \text{with } D(x_i, x_j) \leq d\}$$

All face pairs in test set that belong to same identities

- False accepts: set of face pairs incorrectly classified as same for threshold *d*

$$\mathrm{FA}(d) = \{(i,j) \in \mathcal{P}_{\text{diff}}, \text{with } D(x_i, x_j) \leq d\}$$

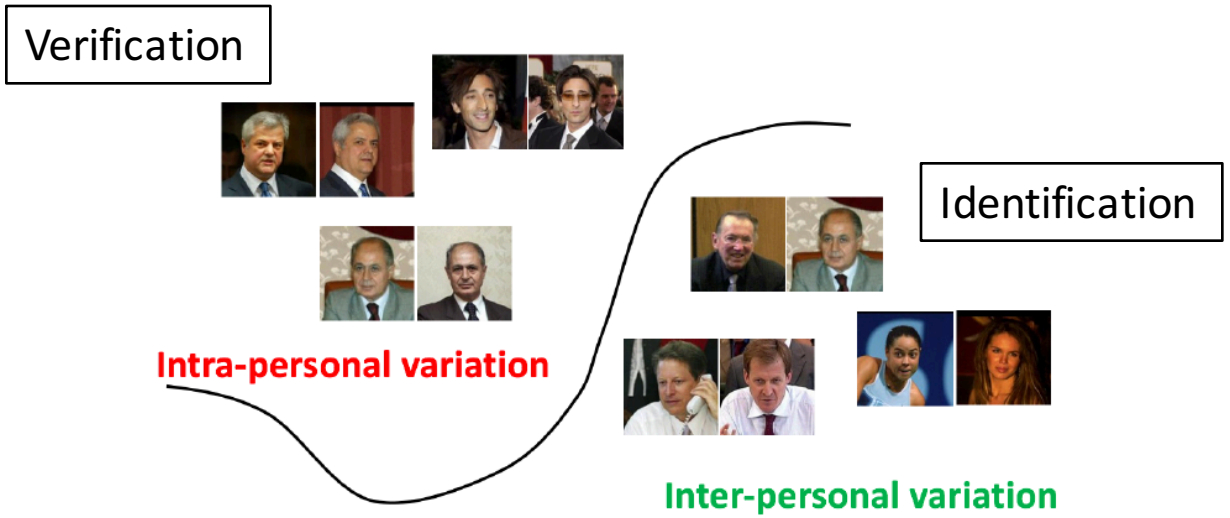
All face pairs in test set that belong to different identities

- Determine validation rate and false accept rate:

$$\mathrm{VAL}(d) = \frac{|\mathrm{TA}(d)|}{|\mathcal{P}_{\text{same}}|}, \quad \mathrm{FAR}(d) = \frac{|\mathrm{FA}(d)|}{|\mathcal{P}_{\text{diff}}|}$$

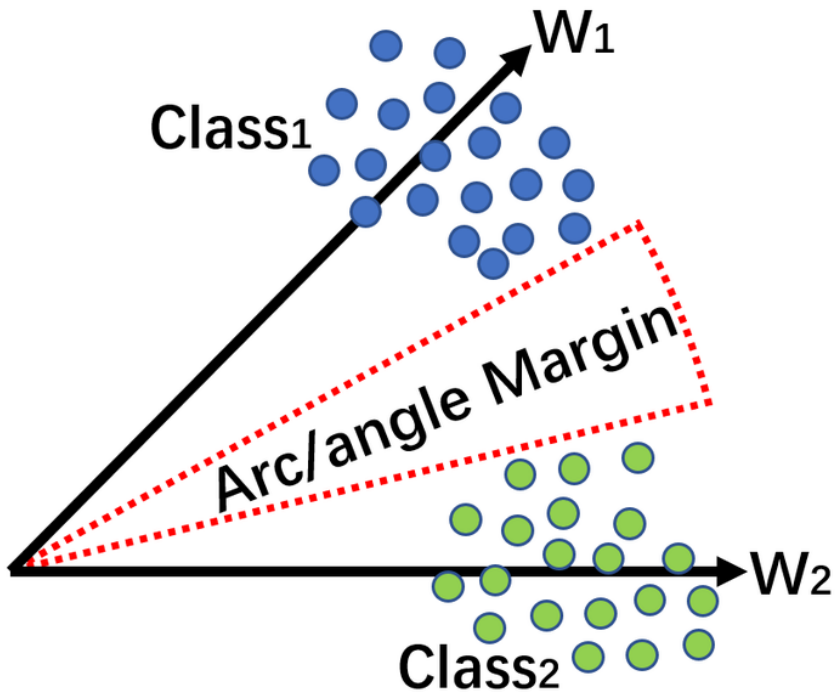
Some Recipes for Face Recognition to Work

- Balance between identification and verification signals



Some Recipes for Face Recognition to Work

- Balance between identification and verification signals
- Softmax easier to train, metric learning for large number of identities
- Best objectives use some combination of both



Some Recipes for Face Recognition to Work

- Balance between identification and verification signals
- Softmax easier to train, metric learning for large number of identities
- Best objectives use some combination of both
- Large-scale data is essential for highly discriminative performance
- Clean and noise-free data leads to large improvements

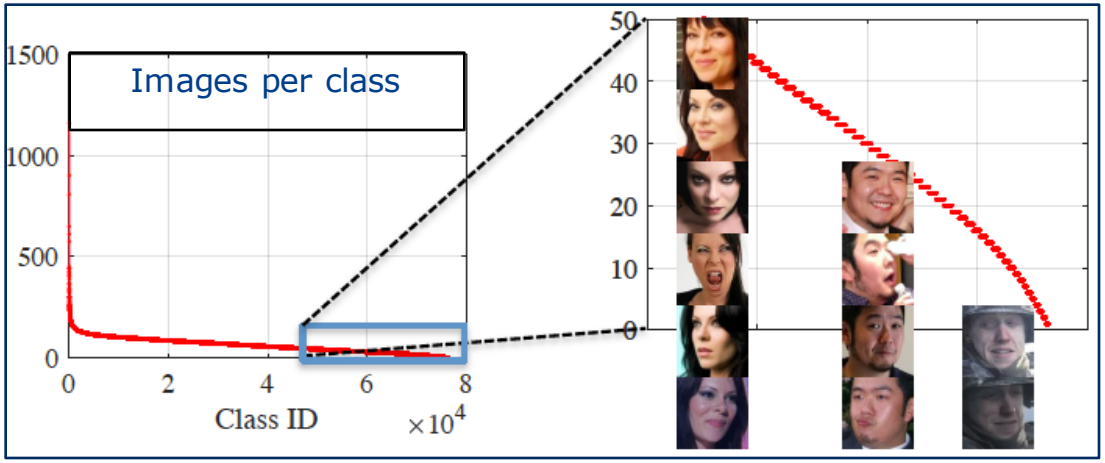
Dataset	#Identities	#Images	Source	Cleaned?	Availablity
LFW [7]	5K	13K	Search Engine	Automatic Detection	Public
CelebFaces [19,20]	10K	202K	Search Engine	Manually Cleaned	Public
VGG-Face [15]	2.6K	2.5M	Search Engine	Semi-automated Clean	Public
CASIA-WebFace [25]	10k	0.5M	IMDb	Automatic Clean	Public
MS-Celeb-1M(v1) [5]	100k	10M	Search Engine	None	Public
MegaFace [13]	670K	4.7M	Flickr	Automatic Cleaned	Public
Facebook [21]	4k	4.4M	–	–	Private
Google [18]	8M	200M	–	–	Private
IMDb-Face	**59K**	**1.7M**	**IMDb**	**Manually Cleaned**	**Public**

Dataset	#Iden.	#Imgs.	Rank-1 (%)		
			Softmax	Center Loss	A-Softmax
CelebFaces	10k	0.20M	36.15	42.54	43.72
CASIA-WebFace	10.5k	0.49M	65.17	68.09	70.89
MS-Celeb-1M(V1)	96k	8.6M	71.70	73.82	73.99
MegaFace	670k	4.7M	64.32	64.71	66.95
IMDbFace	59k	1.7M	**74.75**	**79.41**	**84.06**

[Wang et al., IMDbFace, ECCV 2018]

Some Recipes for Face Recognition to Work

- Balance between identification and verification signals
- Softmax easier to train, metric learning for large number of identities
- Best objectives use some combination of both
- Large-scale data is essential for highly discriminative performance
- Clean and noise-free data leads to large improvements
- Even largest datasets need to be augmented for coverage
- In general, data augmentation better than input normalization



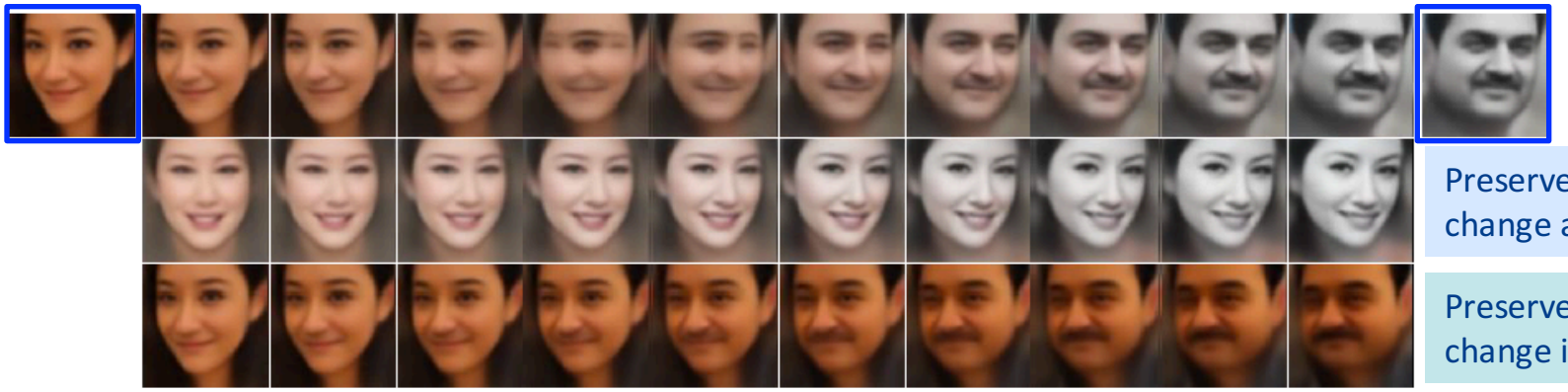
Few classes have many examples, most classes have few examples.

- Learn distribution of regular classes (mean and variance)
- Transfer to low-shot classes to obtain diverse new examples.

[Yin et al., Long-Tail Face Recognition, CVPR 2019]

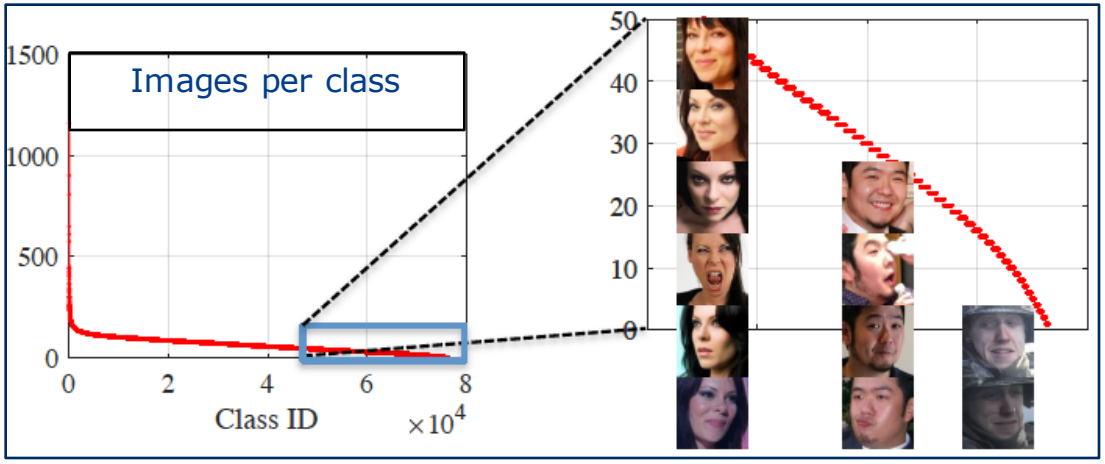
Some Recipes for Face Recognition to Work

Row 1: feature interpolation, Rows 2 and 3: feature transfer



Preserve identity, change attribute

Preserve attribute, change identity



Few classes have many examples, most classes have few examples.

- Learn distribution of regular classes (mean and variance)
- Transfer to low-shot classes to obtain diverse new examples.

[Yin et al., Long-Tail Face Recognition, CVPR 2019]

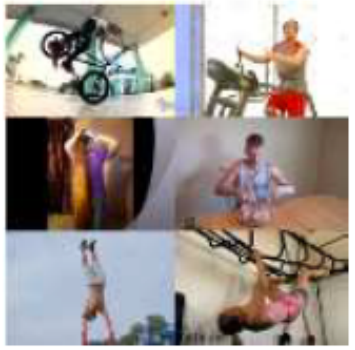
Some Recipes for Face Recognition to Work

- Balance between identification and verification signals
- Softmax easier to train, metric learning for large number of identities
- Best objectives use some combination of both
- Large-scale data is essential for highly discriminative performance
- Clean and noise-free data leads to large improvements
- Even largest datasets need to be augmented for coverage
- In general, data augmentation better than input normalization
- Large batch sizes for state-of-the-art performance (multi-node training)

Human Pose Estimation

Human Pose: Challenges

- Many applications, but challenging



Action recognition

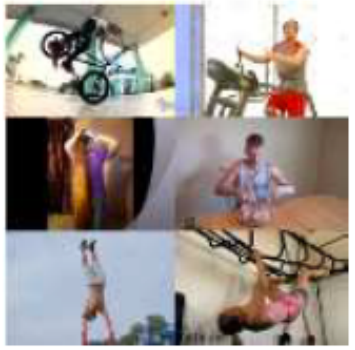
Human Parsing

Game / Animation

Human Pose: Challenges

- Many applications, but challenging

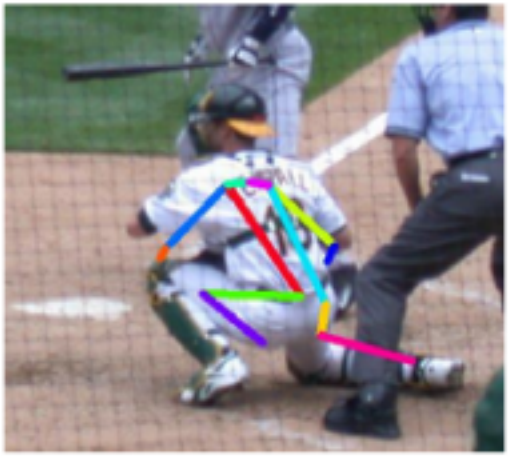
- Strong articulations
- Small and barely visible joints
- Occlusions
- The need to capture context



Action recognition

Human Parsing

Game / Animation

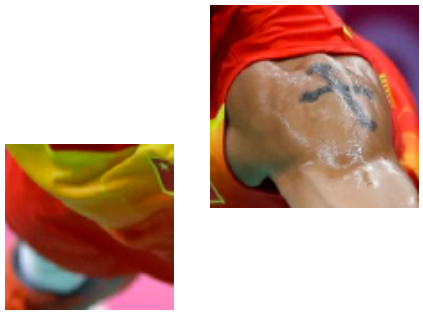


Local and Global Needed for Pose Estimation

- Local part-based detectors are not sufficient for pose estimation

Local and Global Needed for Pose Estimation

- Local part-based detectors are not sufficient for pose estimation



Local and Global Needed for Pose Estimation

- Local part-based detectors are not sufficient for pose estimation

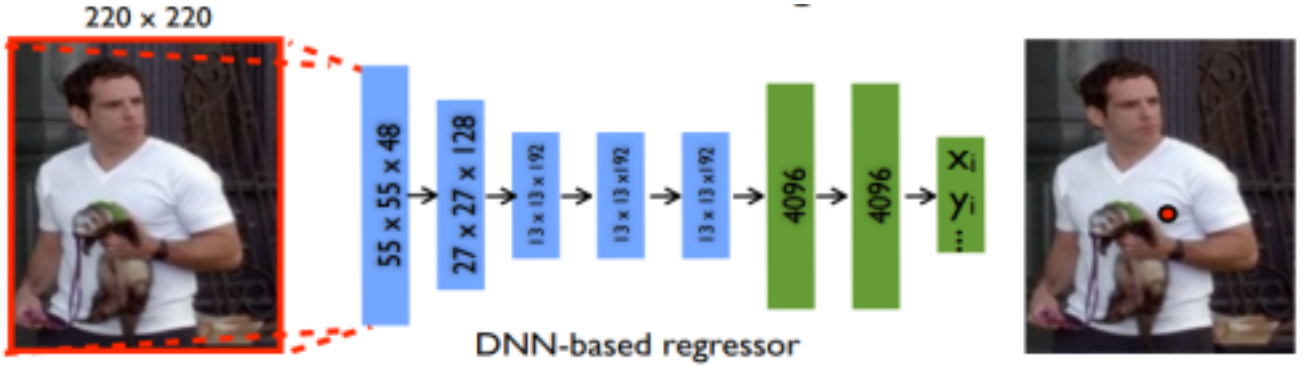
Deep Networks for Holistic Pose Estimation

- Local part-based detectors are not sufficient for pose estimation

- Deep network takes whole image as input
- Global information from whole image used for each joint regression

- Regresses joint locations, no need to design detectors

- No need to explicitly model joint relations and spatial constraints



[Toshev and Szegedy, DeepPose]

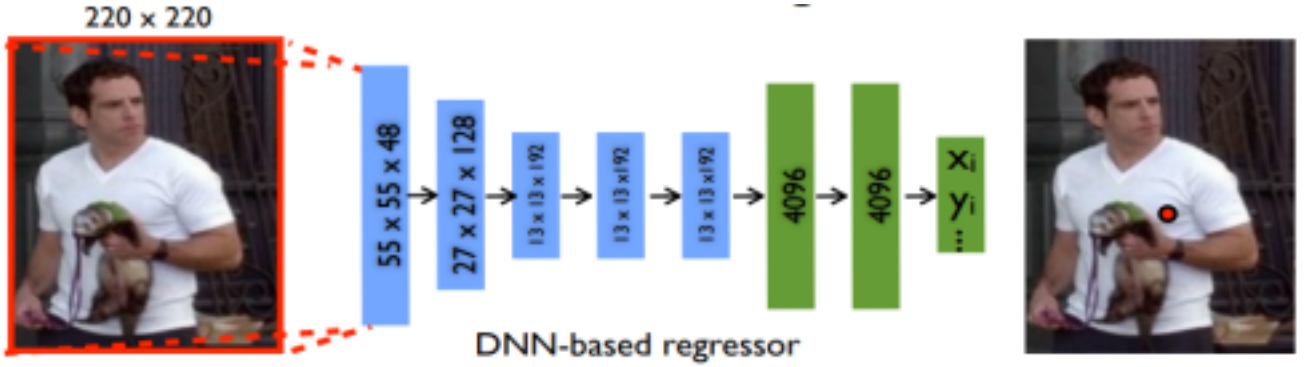
Deep Networks for Holistic Pose Estimation

- Input: image *x*
- CNN learns prediction functions $\psi_i(x;\theta)$
- Output: normalized pose vector $y^* = N^{-1}(\psi(N(x);\theta))$
- Normalization is with respect to bounding box

$$\mathbf{y} = (\ldots, \mathbf{y}_i^T, \ldots)^T, i \in \{1, \ldots, k\} \qquad N(\mathbf{y}_i; b) = \begin{pmatrix} 1/b_w & 0 \\ 0 & 1/b_h \end{pmatrix} (\mathbf{y}_i - b_c)$$

- Training: minimize L2 distance between prediction and ground truth pose

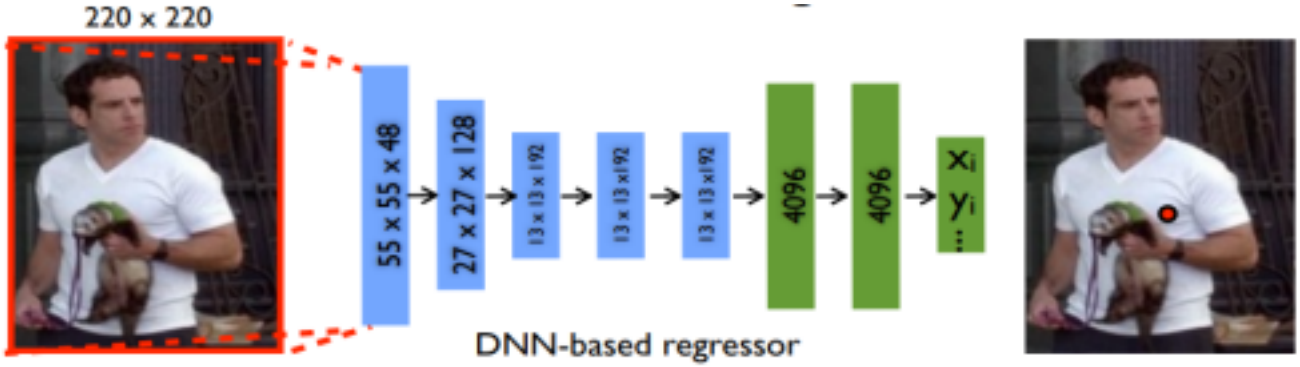
$$\arg\min_{\theta} \sum_{(x,y)\in D_N} \sum_{i=1}^{k} ||\mathbf{y}_i - \psi_i(x;\theta)||_2^2$$



Deep Networks for Holistic Pose Estimation

- Advantages:
  - Simple, yet holistic
  - No need to define losses that capture interactions
  - Instead, all hidden layers are shared by joint regressors

- Disadvantages:
  - Limited ability to consider details

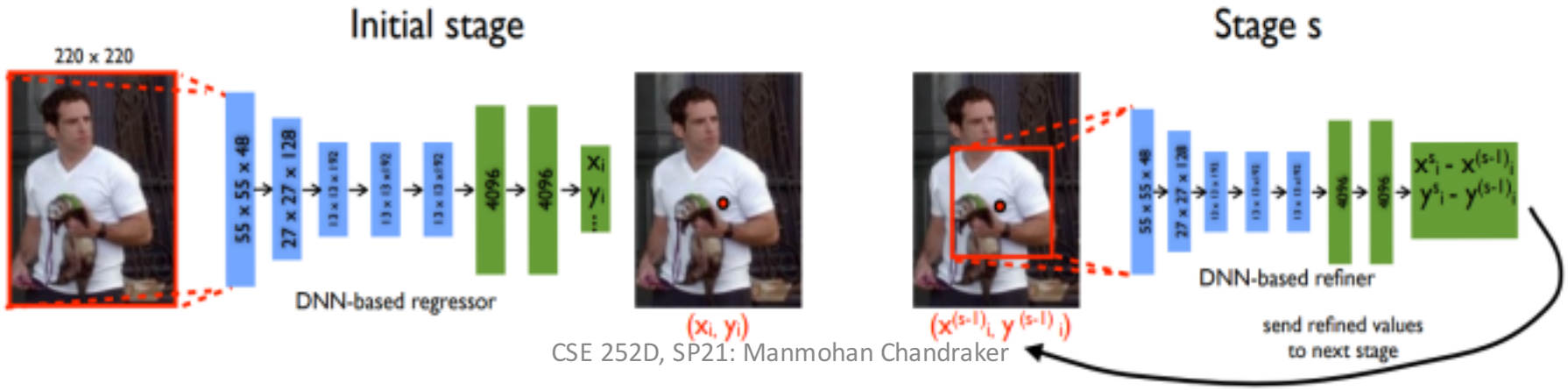


Cascade of Regressors

- Same network architecture but different parameters (that is, θ_s)
- At stage *s* > 1, CNN regressor trained with predicted joint from stage *s*-1

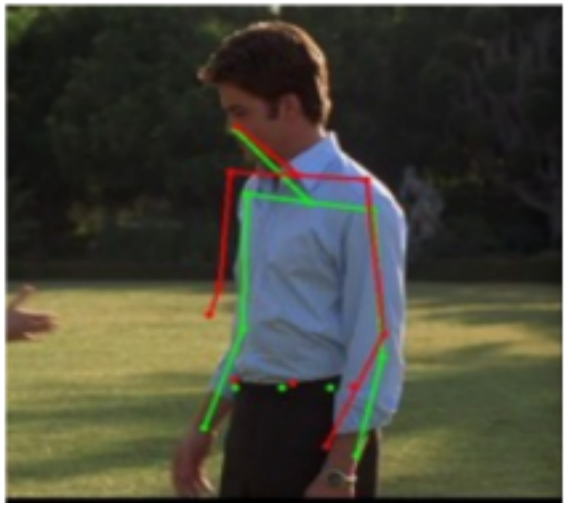
$$\text{Stage 1}: \quad \mathbf{y}^1 \leftarrow N^{-1}(\psi(N(x; b^0); \theta_1); b^0)$$

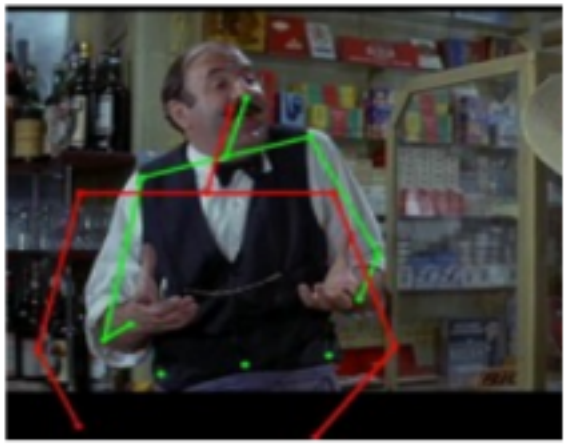
$$\text{Stage } s: \quad \mathbf{y}_i^s \leftarrow \mathbf{y}_i^{(s-1)} + N^{-1}(\psi_i(N(x; b); \theta_s);$$
$$\text{for } b = b_i^{(s-1)}$$
$$b_i^s \leftarrow (\mathbf{y}_i^s, \sigma\text{diam}(\mathbf{y}^s), \sigma\text{diam}(\mathbf{y}^s))$$



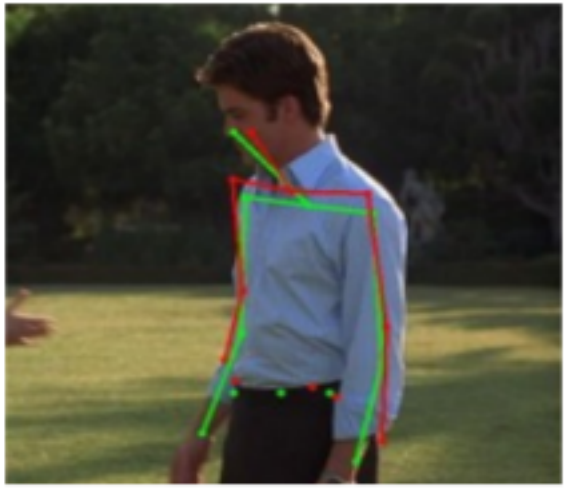
Cascade of Regressors

Stage 0

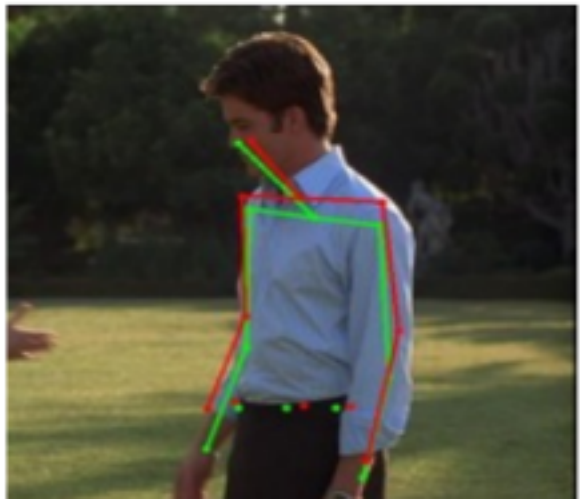


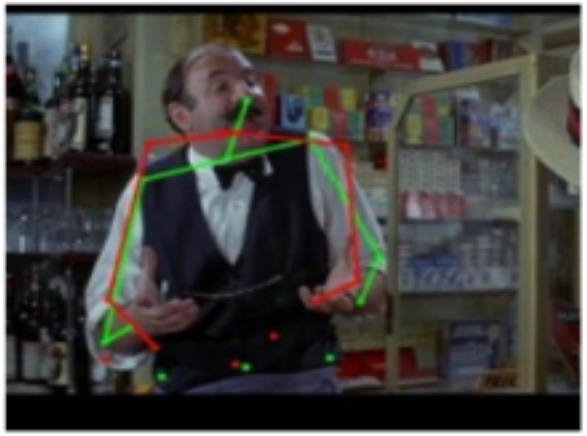


Stage 1



Stage 2

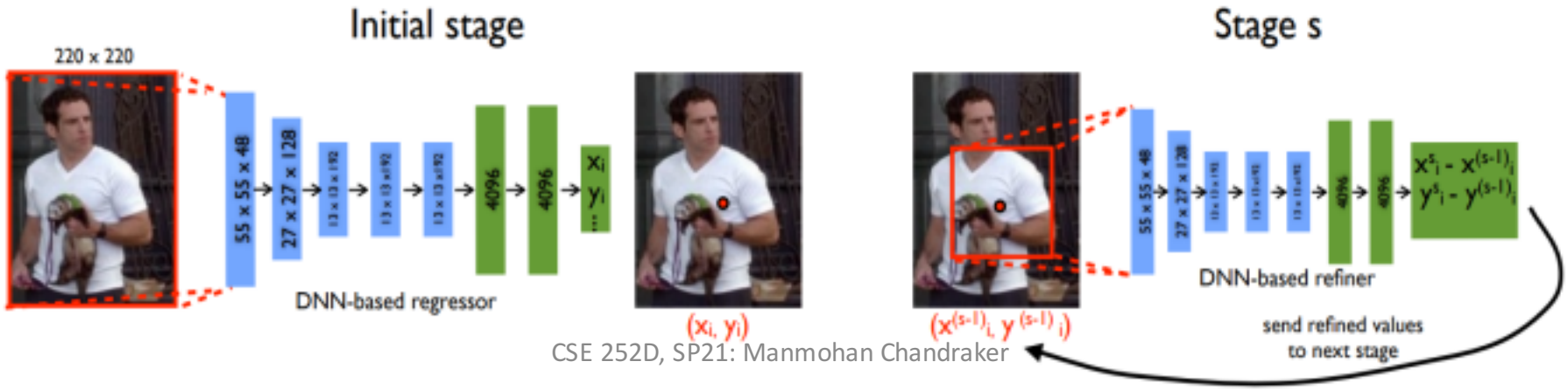




Cascade of Regressors

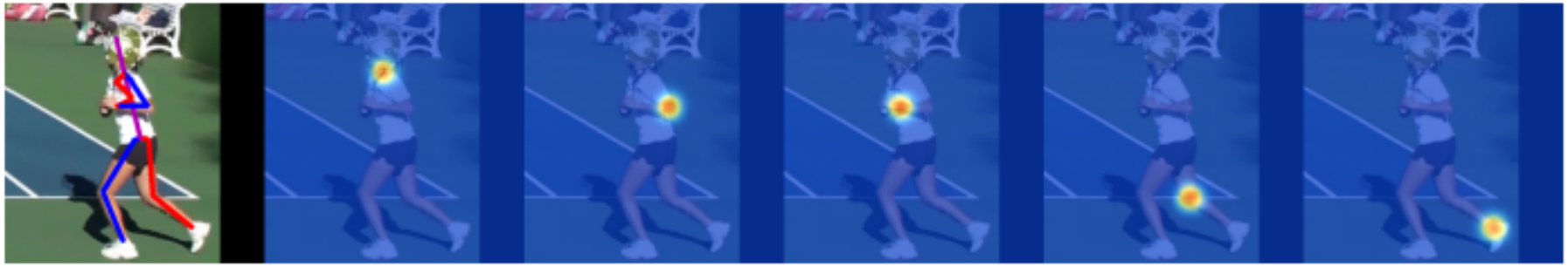
- Advantages:
  - Simple, yet holistic
  - No need to define losses that capture interactions
  - Instead, all hidden layers are shared by joint regressors
  - Increasingly detailed prediction along cascade stages

- Disadvantages:
  - ~~Limited ability to consider details~~
  - One prediction per image, no candidates
  - Depends on quality of initial prediction



Cascade of Heat Maps

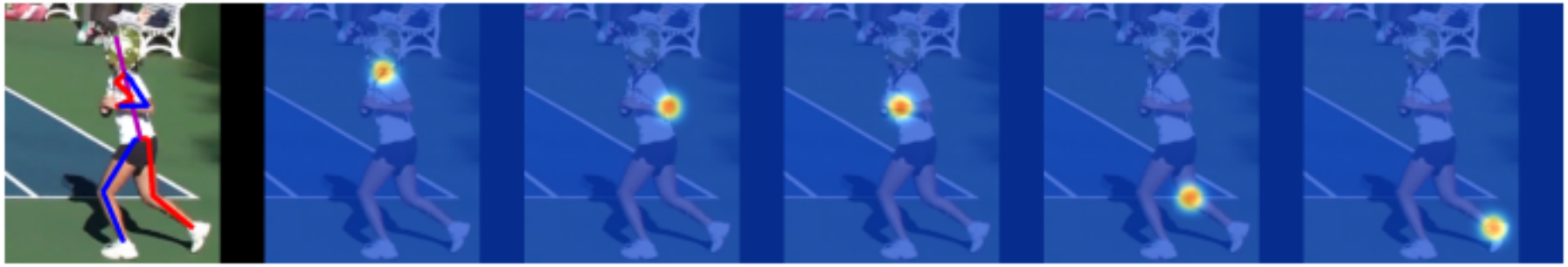
- Predict heat maps instead of pixel locations



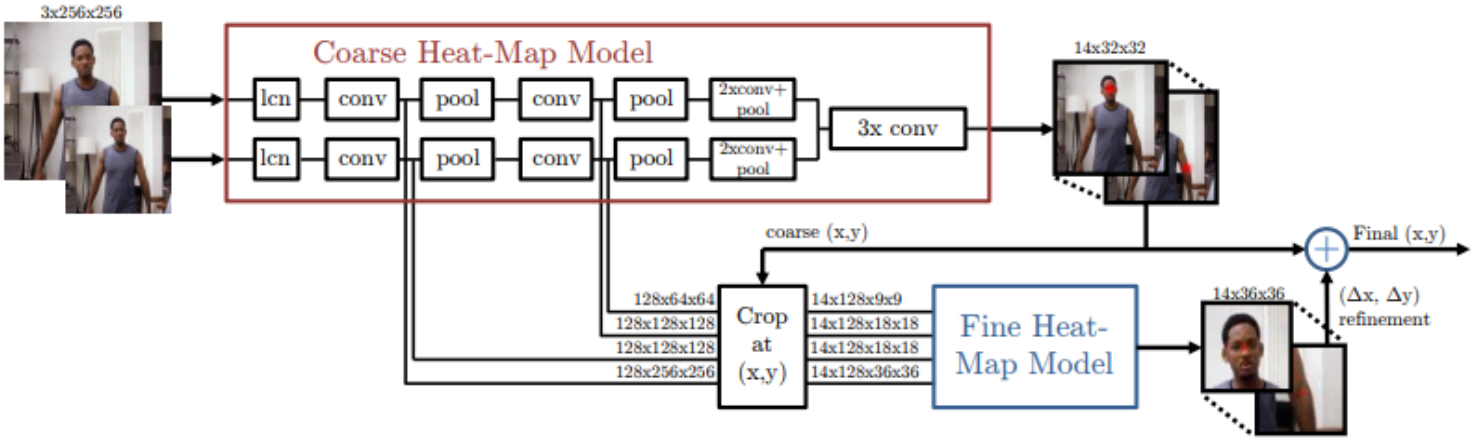
[Tompson et al., Efficient Object Localization]

Cascade of Heat Maps

- Predict heat maps instead of pixel locations



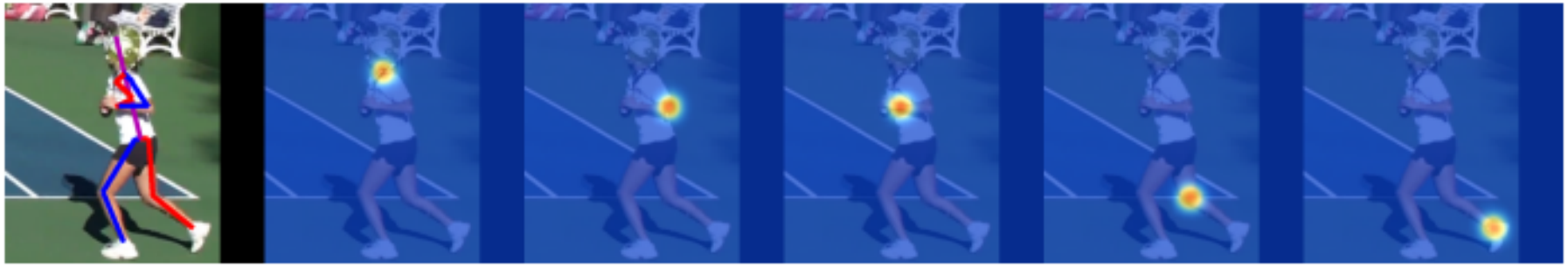
- Input images at multiple scales for multiresolution feature extraction
- Reuse convolutional features across scales



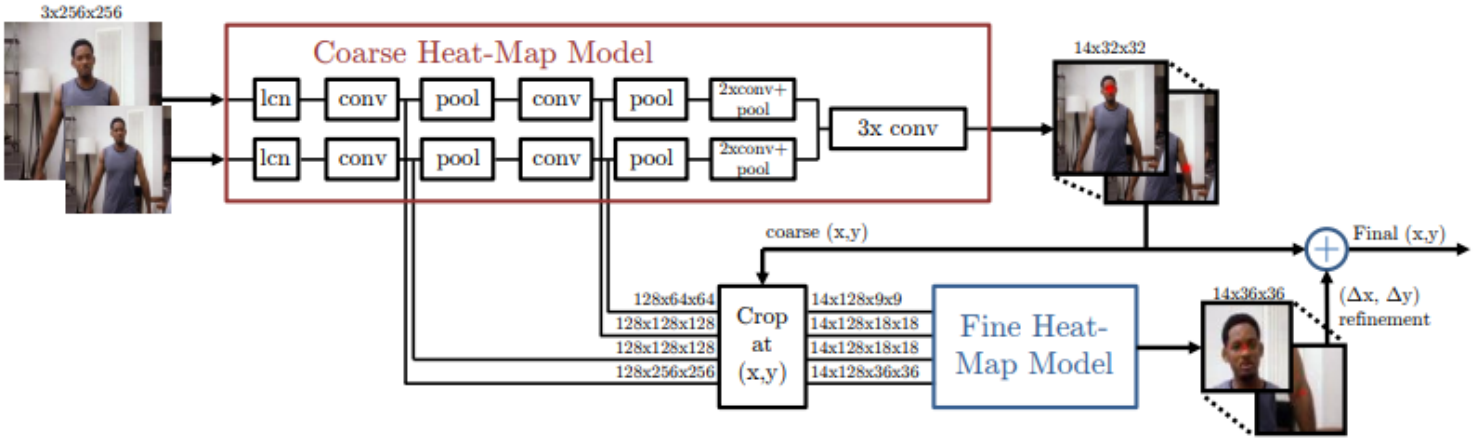
[Tompson et al., Efficient Object Localization]

Cascade of Heat Maps

- Predict heat maps instead of pixel locations



- Input images at multiple scales for multiresolution feature extraction
- Reuse convolutional features across scales



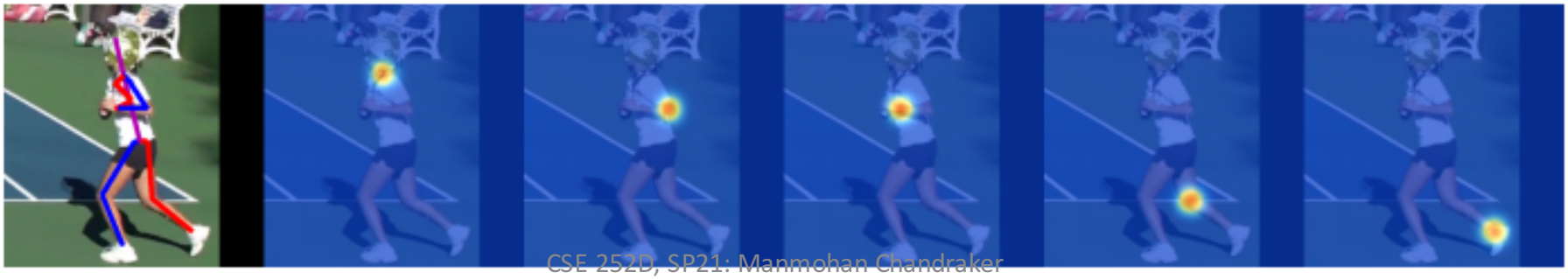
- Minimize MSE between predicted and target heat maps
- Target heat map : apply Gaussian of standard deviation 1.5 pixels

[Tompson et al., Efficient Object Localization]

Cascade of Heat Maps

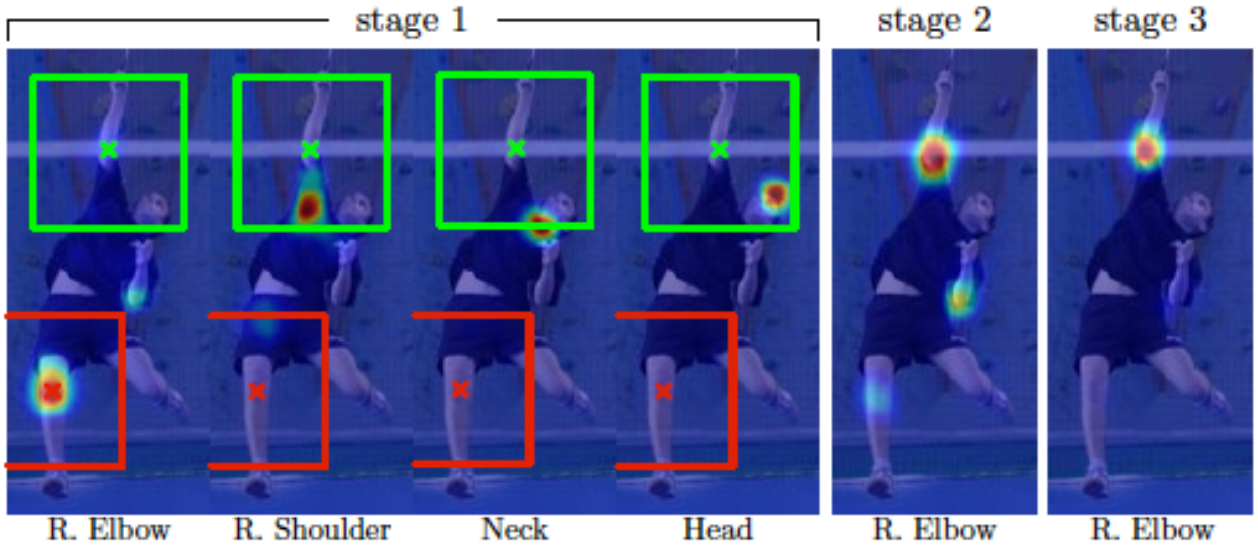
- Advantages:
  - Simple, yet holistic
  - No need to define losses that capture interactions
  - Instead, all hidden layers are shared by joint regressors
  - Increasingly detailed prediction along cascade stages
  - Encodes probability of joint appearing at a pixel

- Disadvantages:
  - ~~Limited ability to consider details~~
  - ~~One prediction per image, no candidates~~
  - Depends on quality of initial prediction
  - Not enough modeling of spatial structure
  - Cannot reason about occluded parts, or those outside window



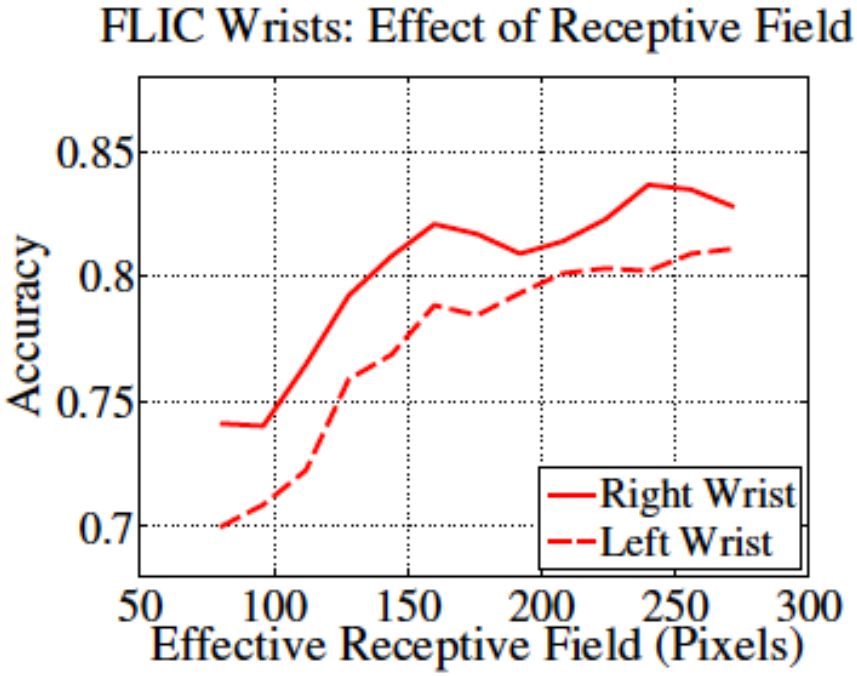
Spatial Structure Helps Localization

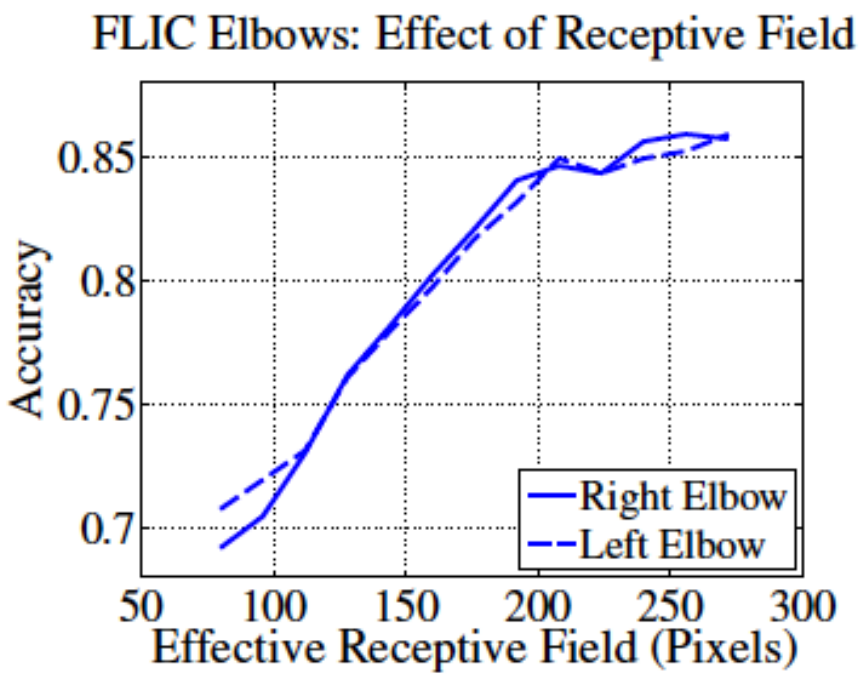
- Shoulder, neck and head localization correct error in location for elbow



Spatial Structure Helps Localization

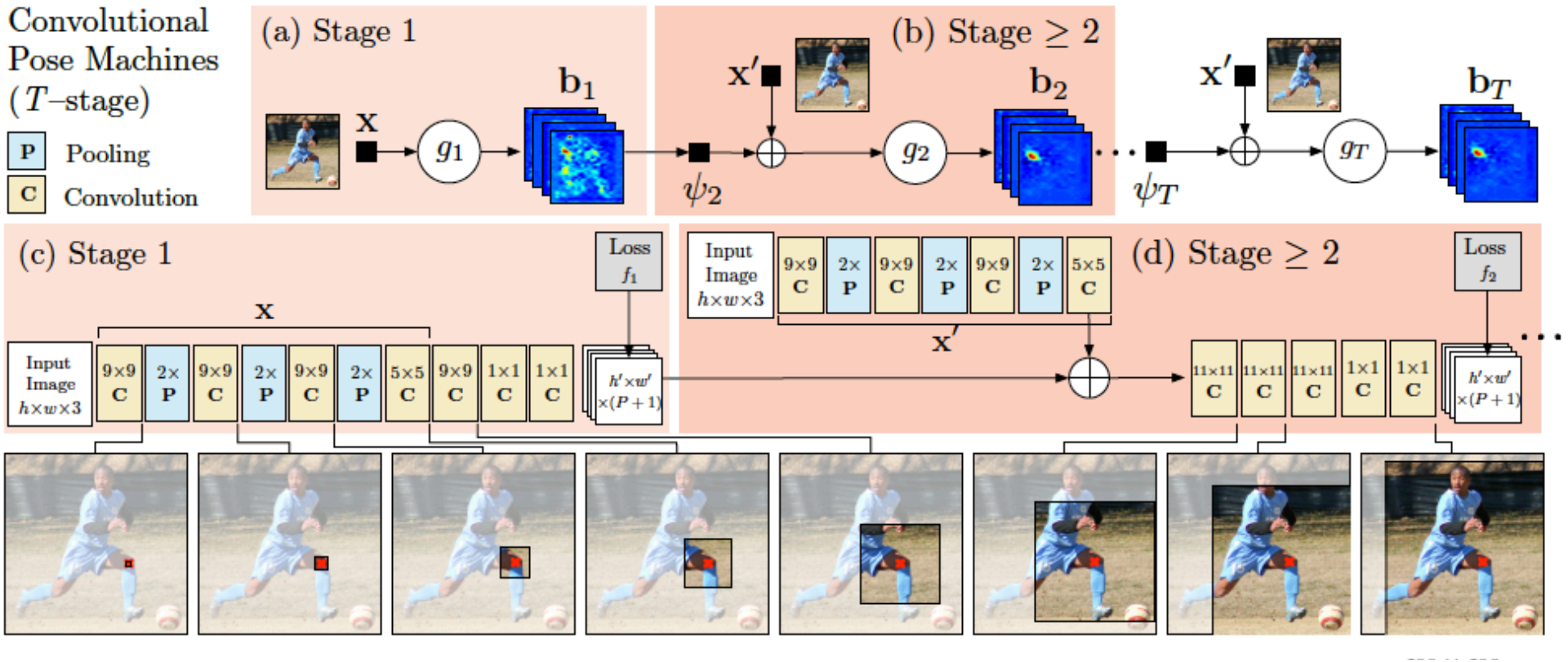
- Goal: achieve large receptive fields to learn complex and long-range interactions





Convolutional Pose Machines

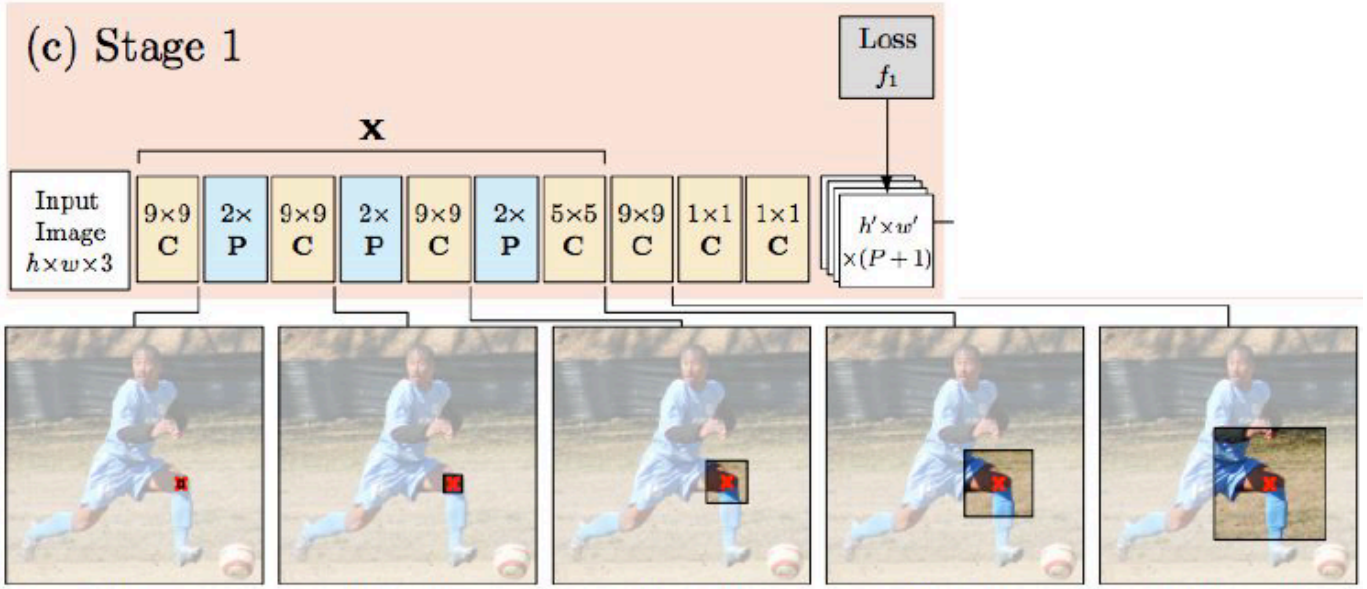
- Goal: achieve large receptive fields to learn complex and long-range interactions
- Each stage sees two inputs
  - Image features
  - Context features based on output of previous stage



[Wei et al., Convolutional Pose Machines]

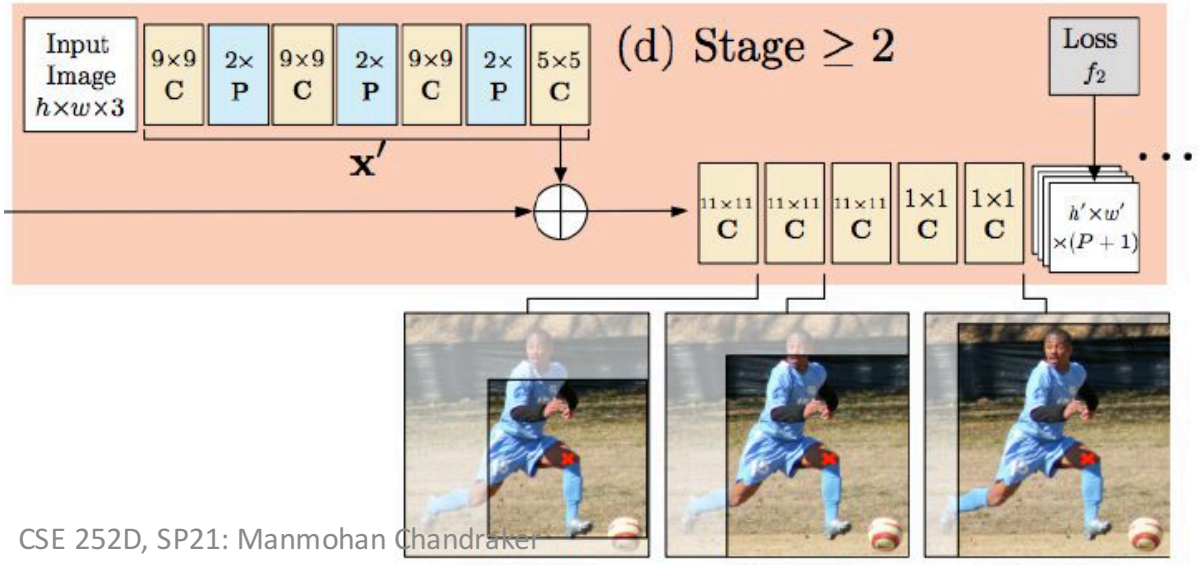
Convolutional Pose Machines: Stage 1

- Predict part belief largely based on local image values
- Output P+1 heat maps (P parts and 1 background)
- Small receptive field
  - Captures relation between head and shoulders, but not head and knees



Convolutional Pose Machines: Next Stages

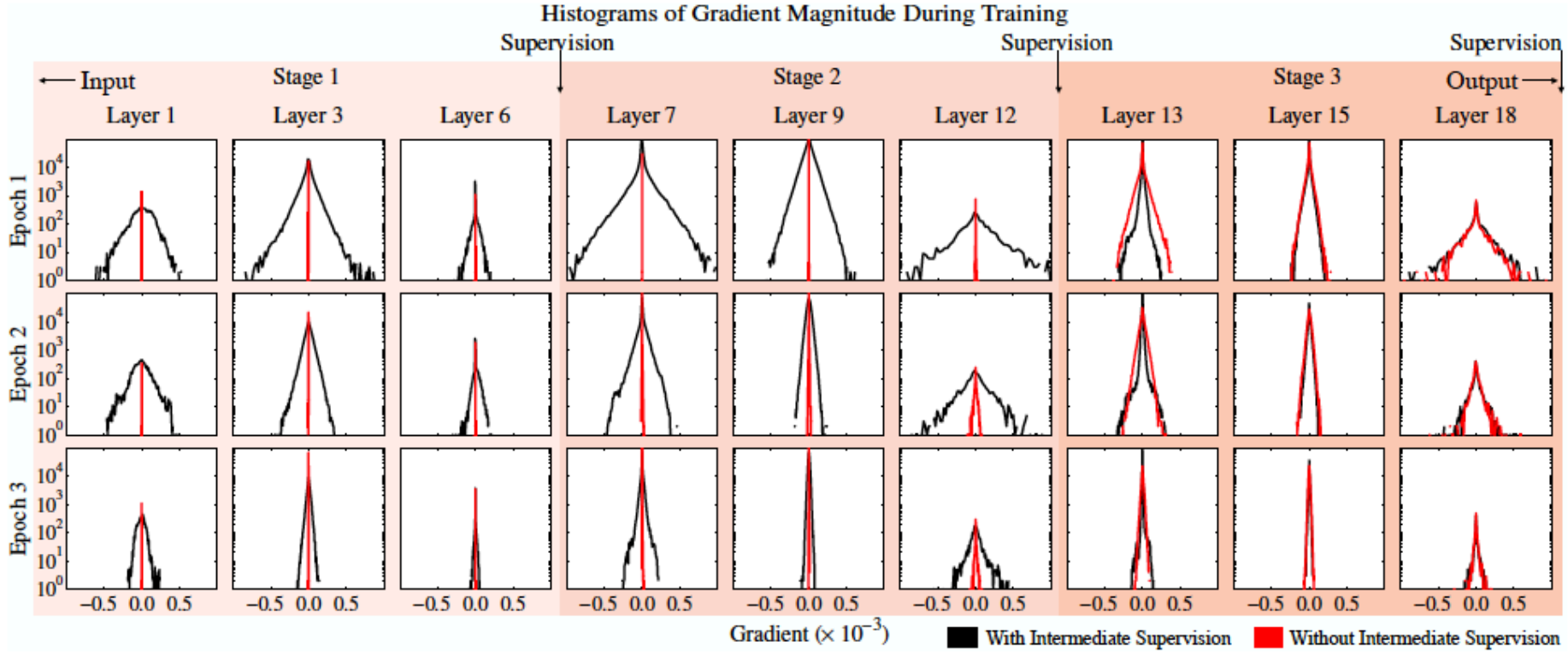
- Image features **x'** from the previous stage
- Context function ψ encodes landscape of belief maps around part locations
  - In practice, ψ is just the receptive field
  - Network decides how to combine features and learn higher relations
  - Previously: hand-defined potential functions in graphical model
- Three ways to increase size of receptive field
  - More pooling: lose local details
  - Larger filters: increase number of parameters
  - More layers: vanishing gradients



Convolutional Pose Machines: Gradients

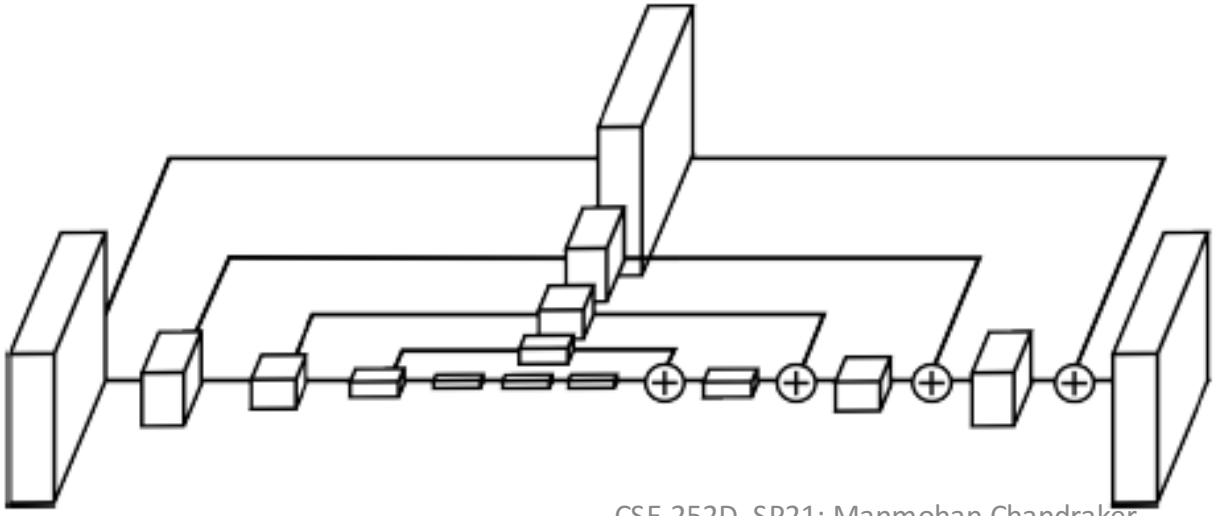
- Magnitude of backpropagated gradients decreases rapidly in initial layers
- Use intermediate supervision to ensure greater variance in gradients

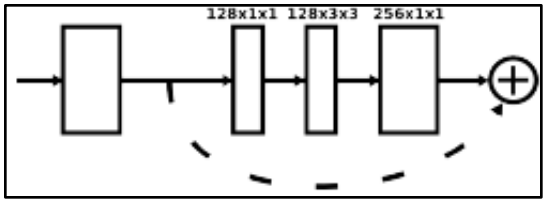
$$f_t = \sum_{p=1}^{P+1} \sum_{z \in \mathcal{Z}} \| b_t^p(z) - b_*^p(z) \|_2^2 \ , \quad \mathcal{F} = \sum_{t=1}^{T} f_t.$$



Hourglass Network

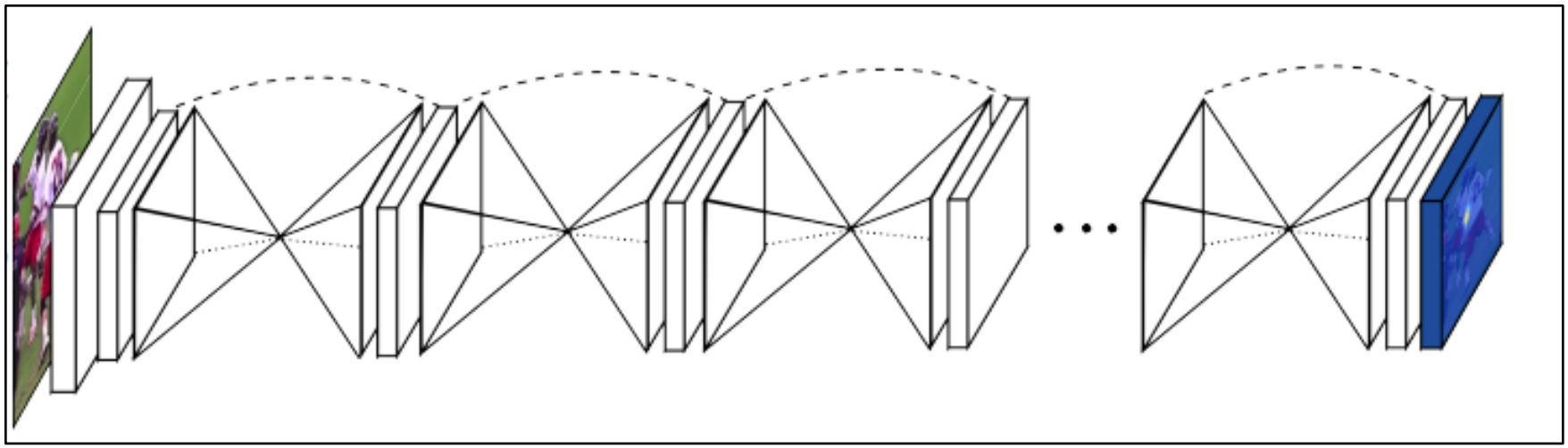
- Local appearance: needed for accurate part detection
- Global reasoning: orientation of body, limb arrangement, part relationships
- Simple design to process multiple scales and achieve pixel-wise predictions
- Bottom-up reasoning: Convolution and max-pooling to very low 4x4 resolution
- Top-down reasoning: Upsample and combine with skip connection





Stacked Hourglass Network

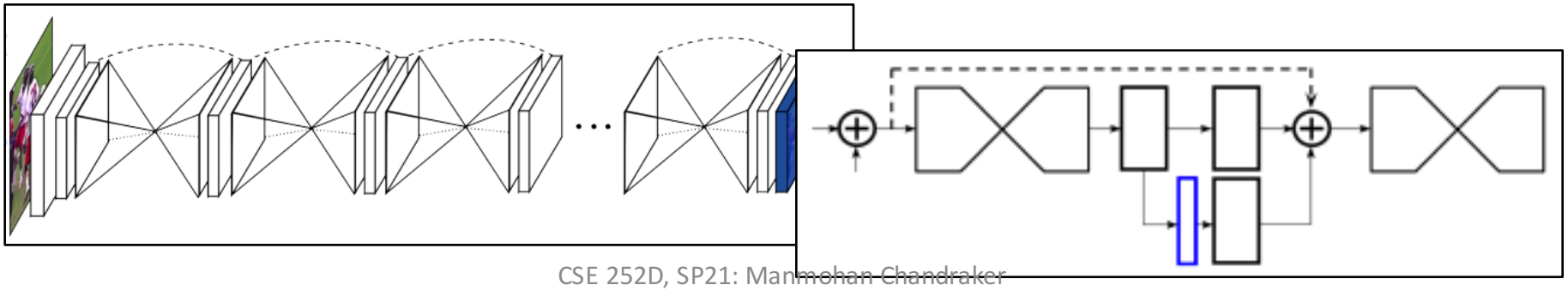
- Multiple iterative stages allow refinement
- Each stage does full bottom-up and top-down processing (no weights shared)



[Newell et al., Stacked Hourglass Networks]

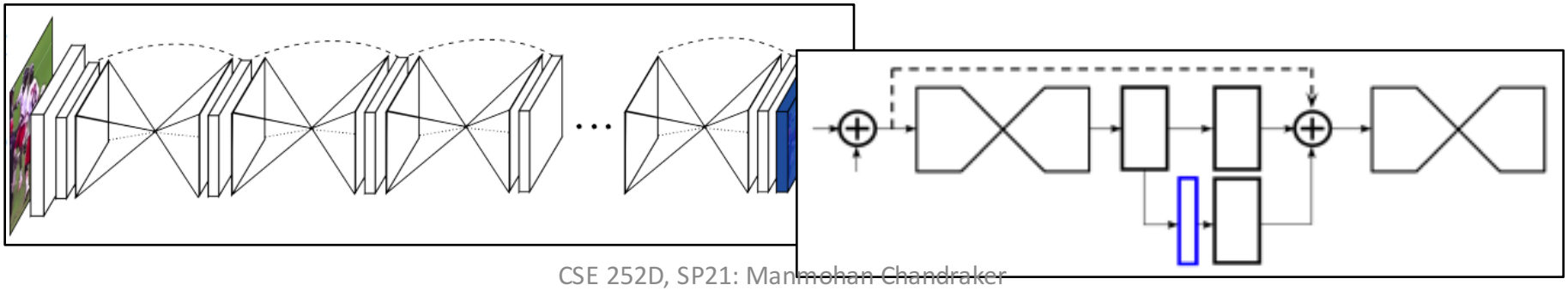
Stacked Hourglass Network

- Multiple iterative stages allow refinement
- Each stage does full bottom-up and top-down processing (no weights shared)
- Can apply intermediate supervision for each stage
  - Network has had a chance to reason both locally and globally
  - Subsequent hourglass modules can reassess high-order spatial relations
  - Ask network to repeatedly reason across scales
  - 1x1 convolution to add intermediate heatmaps to feature channels



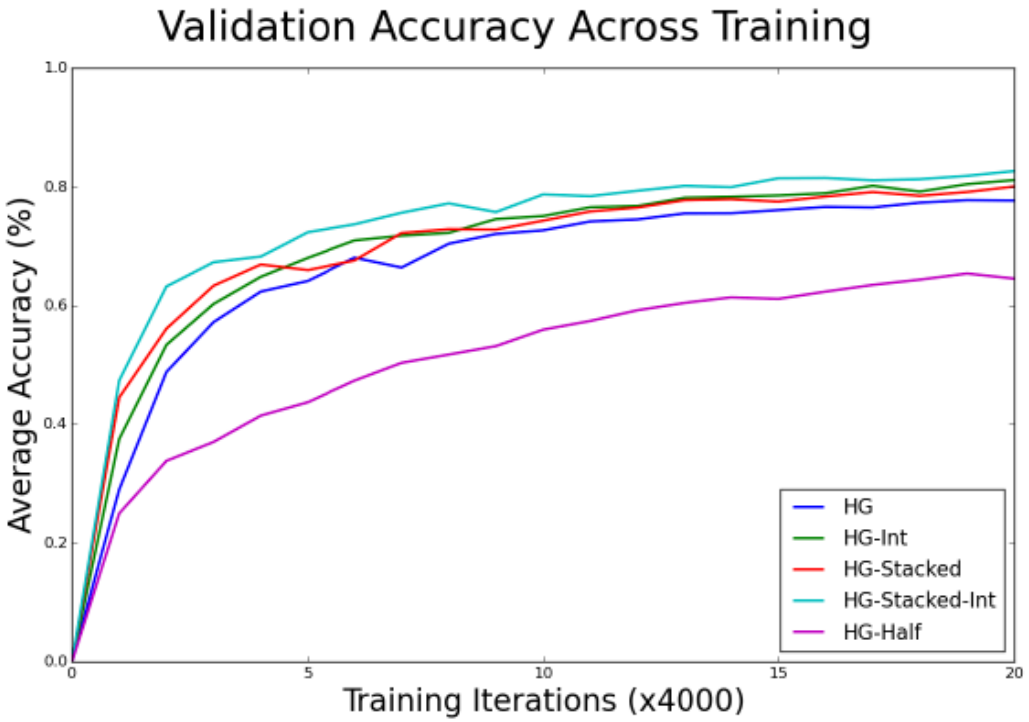
Stacked Hourglass Network

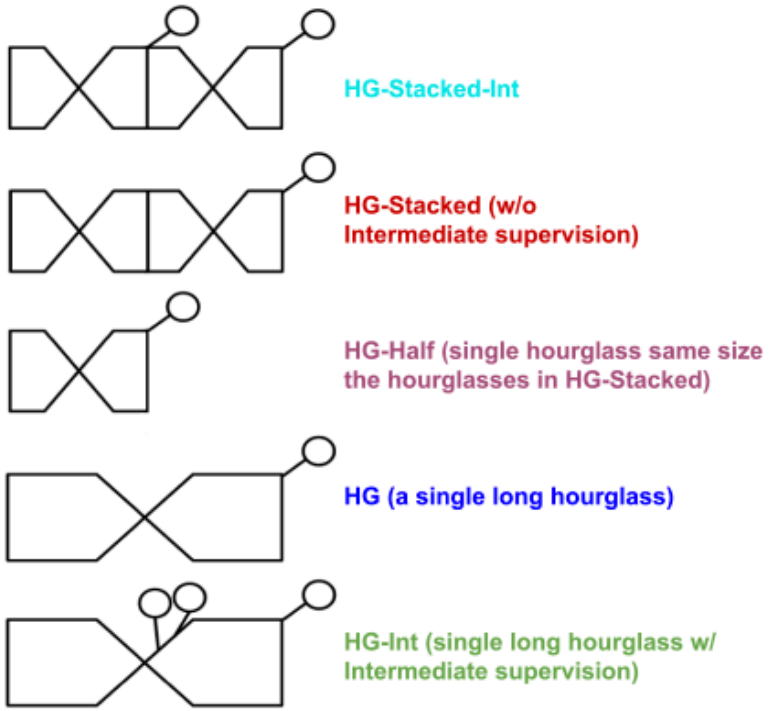
- Multiple iterative stages allow refinement
- Each stage does full bottom-up and top-down processing (no weights shared)
- Can apply intermediate supervision for each stage
  - Network has had a chance to reason both locally and globally
  - Subsequent hourglass modules can reassess high-order spatial relations
  - Ask network to repeatedly reason across scales
  - 1x1 convolution to add intermediate heatmaps to feature channels
- Intermediate supervision not straightforward for single hourglass module
  - Cannot apply before pooling since only local information available
  - High-order features only present at low resolutions
  - After upsampling, have high-order information and higher resolution
  - But cannot re-evaluate features globally relative to each other



Stacked Hourglass Network

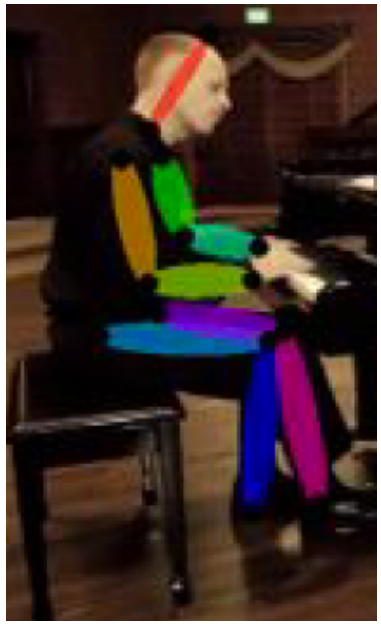
- Depth helps: single long hourglass better than short hourglass
- Cascading better than just depth: stacked hourglass better than long hourglass
- Intermediate supervision helps both single and stacked models





Evaluation of Human Pose Estimation

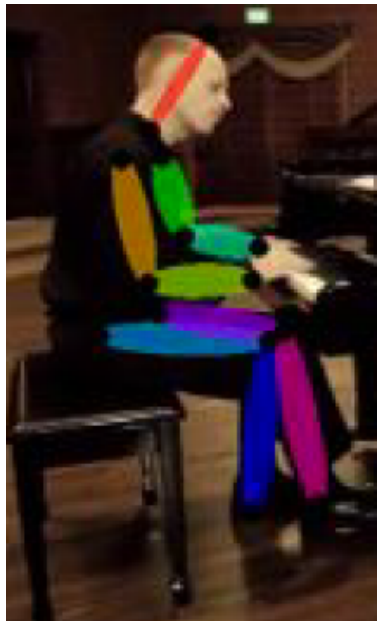
- Percentage of Correct Parts
  - A part (limb) is correct if sum of its joint errors is less than half limb-length
  - Penalizes shorter limbs since they have smaller thresholds



Evaluation of Human Pose Estimation

- Percentage of Correct Parts
  - A part (limb) is correct if sum of its joint errors is less than half limb-length
  - Penalizes shorter limbs since they have smaller thresholds

- Percentage of Correct Keypoints
  - A predicted joint is correct is it lies within a threshold of true location
  - Threshold: 0.2 of torso length, or 0.5 of head bone length
  - Usually shorter limbs have smaller torsos, so does not penalize them

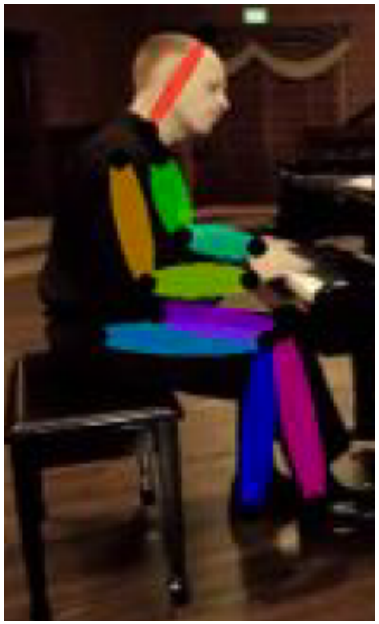


Evaluation of Human Pose Estimation

- Percentage of Correct Parts
  - A part (limb) is correct if sum of its joint errors is less than half limb-length
  - Penalizes shorter limbs since they have smaller thresholds

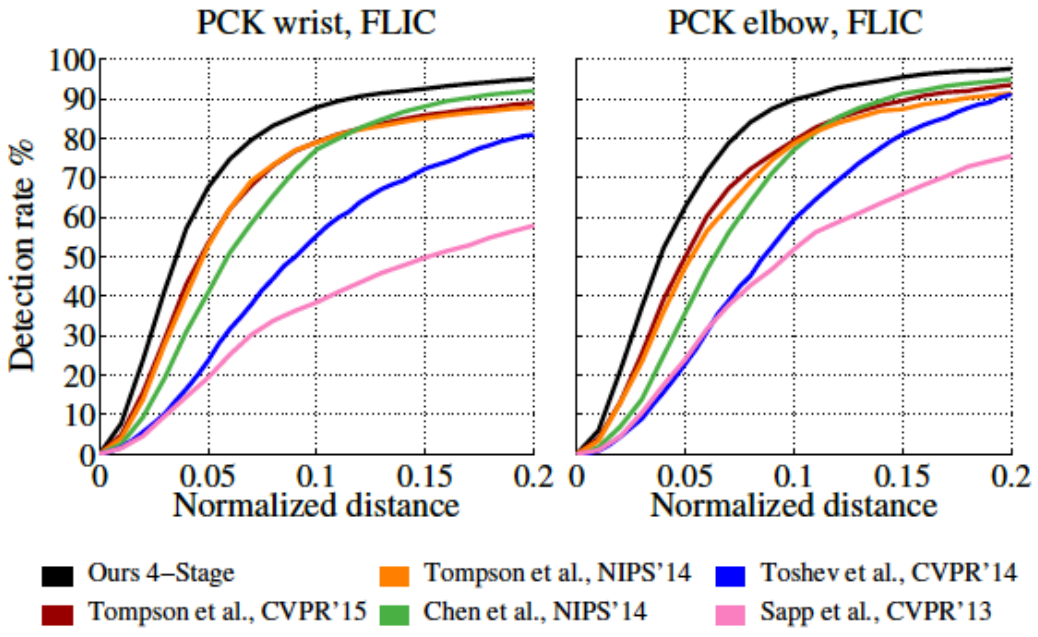
- Percentage of Correct Keypoints
  - A predicted joint is correct is it lies within a threshold of true location
  - Threshold: 0.2 of torso length, or 0.5 of head bone length
  - Usually shorter limbs have smaller torsos, so does not penalize them

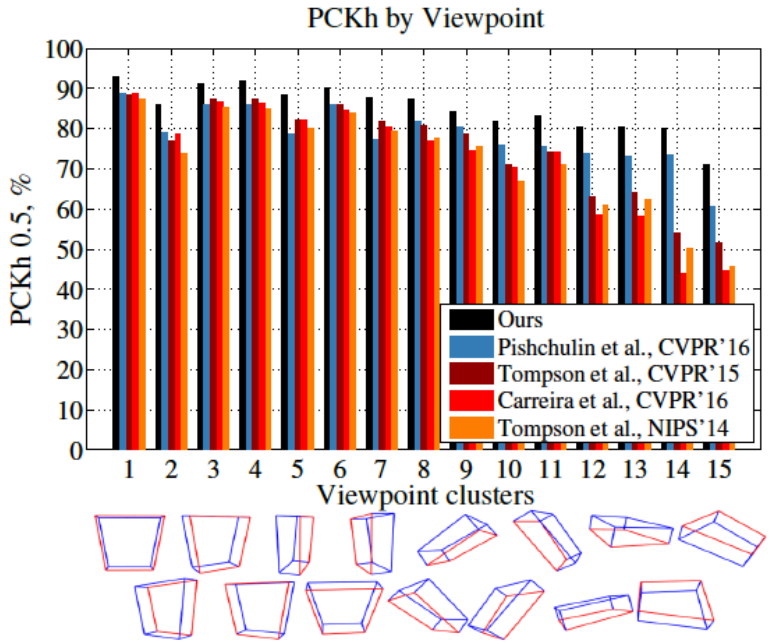
- Object Keypoint Similarity
  - Define OKS = $\frac{\sum_i \exp\left(-d_i^2/2s^2k_i^2\right)\delta\left(v_i > 0\right)}{\sum_i \delta\left(v_i > 0\right)}$
  - d_i: distance between predicted and ground truth keypoint
  - v_i: visibility of keypoint i
  - s: scale of object
  - OKS plays similar role as IoU in detection
  - AP50 : average precision at OKS = 50
  - mAP : mean of AP at OKS = [0.50, 0.55, ..., 0.90, 0.95]



Evaluation of Human Pose Estimation

- Average response from flipped inputs
- Keypoint location quarter offset from highest response to second highest





Convolutional Pose Machines

Recipes for Human Pose Estimation

- Local appearance: needed for accurate part detection
- Global reasoning: orientation of body, limb arrangement, part relationships

Recipes for Human Pose Estimation

- Local appearance: needed for accurate part detection
- Global reasoning: orientation of body, limb arrangement, part relationships

- CNNs are inherently well-suited for this task
  - Multiscale feature extraction
  - Shared hidden layers capture part interactions

- But mechanisms needed to coax local and global performance out of CNNs

Recipes for Human Pose Estimation

- Local appearance: needed for accurate part detection
- Global reasoning: orientation of body, limb arrangement, part relationships

- CNNs are inherently well-suited for this task
  - Multiscale feature extraction
  - Shared hidden layers capture part interactions

- But mechanisms needed to coax local and global performance out of CNNs

- Heat maps
  - Predict probability of joint appearing at a pixel

- Cascades
  - Iteratively refine predictions for fine localization
  - Long-range interactions through wider receptive fields

- Intermediate supervision
  - Prevent vanishing gradients through cascade stages
  - Allow cascade to repeatedly assess local and global information