CVPR 2016

INSTITUT
DES ALGORITHMES
D'APPRENTISSAGE
DE MONTRÉAL

MILA

Université de Montréal

Recombinator Networks

Learning Coarse-to-Fine Feature Aggregation

Sina Honari

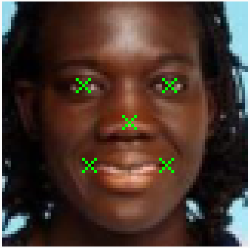
Jason Yosinski

Pascal Vincent

Christopher Pal

Keypoint/Landmark Detection

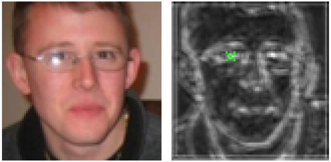
- The problem of localizing important points on images,
  such as eye centers, nose tip, mouth corners



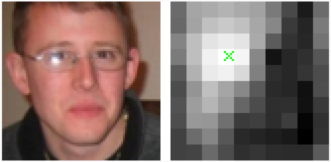
- **Preserving spatial information** is needed for precise keypoint detection.

Motivation

- Convnets are typically composed of **alternating convolutional and max-pooling layers**
- **Network of only convolutional layers:** keeps spatial information, but lots of false positives

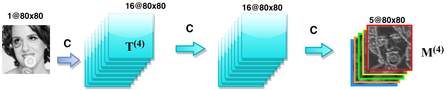


- **Network of convolutional and Max-pooling layers:** gets robust features, but loses precise spatial information



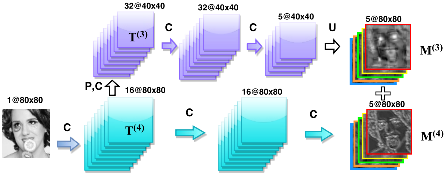
Is there a way to take advantage of robust pooled features <u>and</u> keep spatial information?

Summation-based Networks (SumNet)



- C is a convolutional layer

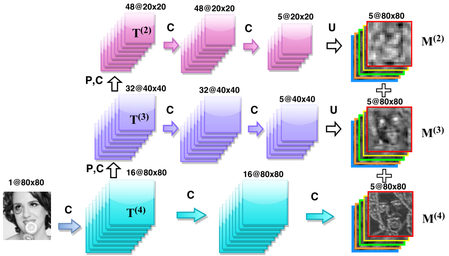
Summation-based Networks (SumNet)



- C is a convolutional layer
- P is a pooling layer
- U is an upsampling layer

- **branch**: horizontal C layers

Summation-based Networks (SumNet)

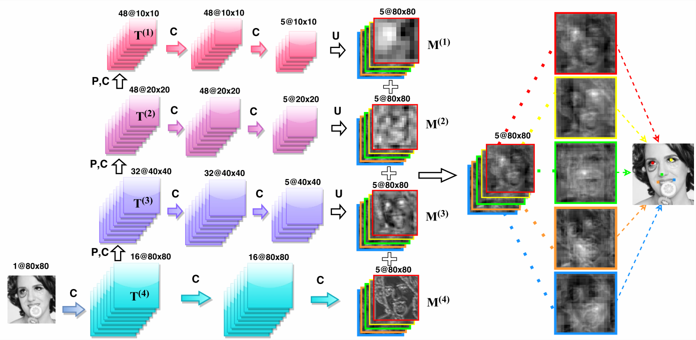


- C is a convolutional layer
- P is a pooling layer
- U is an upsampling layer

- **branch**: horizontal C layers

Summation-based Networks (SumNet)

These models **sum** features of different granularity (FCN[1]/Hypercolumn[2]):



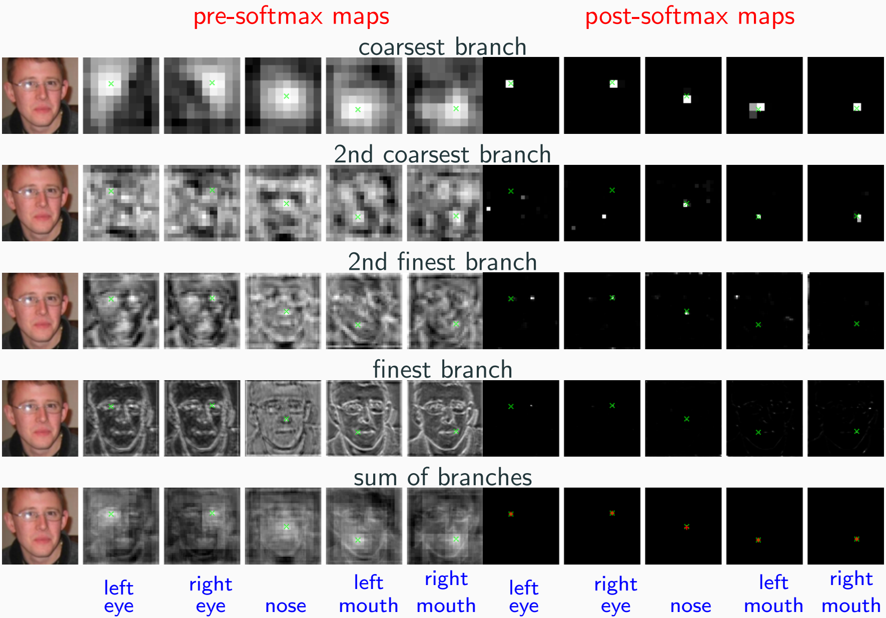
- C is a convolutional layer
- P is a pooling layer
- U is an upsampling layer

- **trunk**: bottom-up C,P layers
- **branch**: horizontal C layers

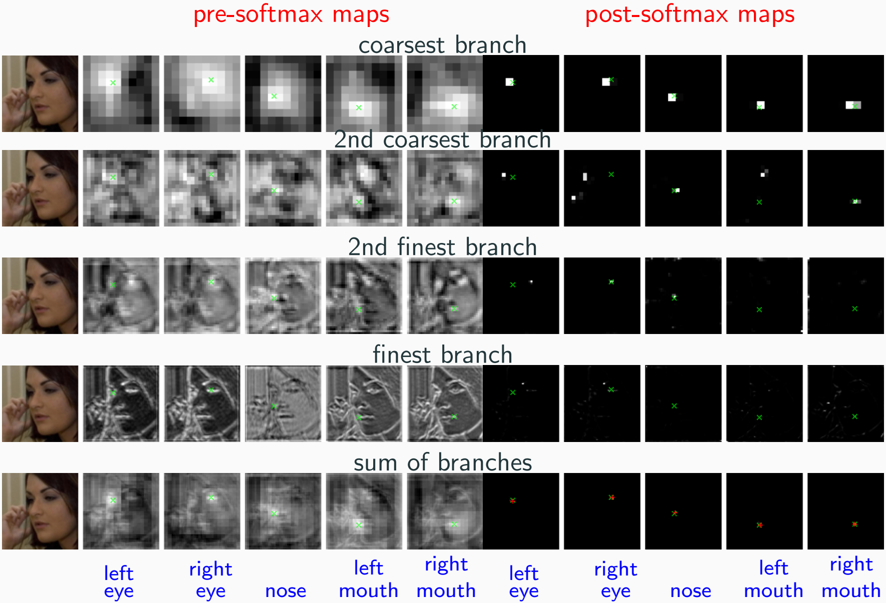
[1] Long, Shelhamer, Darrell. Fully convolutional networks for semantic segmentation. CVPR 2015.

[2] Hariharan, Arbelaez, Girshick, Malik. Hypercolumns for object segmentation and finegrained localization. CVPR 2015.

SumNet Branch Contributions:

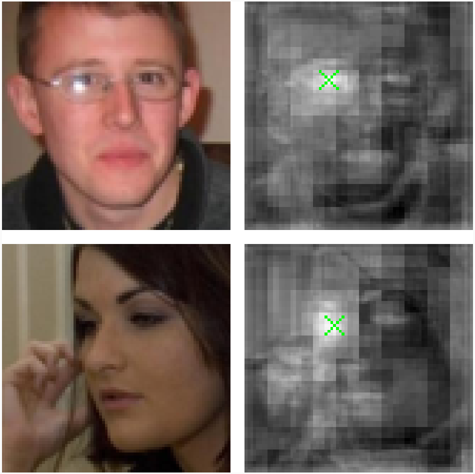


SumNet Branch Contributions:



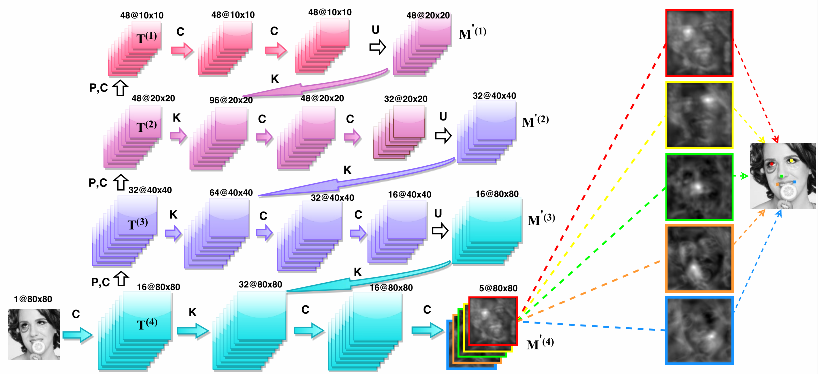
SumNet vs. RCN Pre-Softmax Maps

SumNet



Recombinator Networks (RCN)

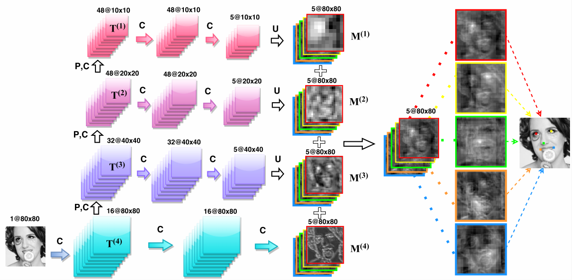
The model feeds coarse features into finer layers early in their computation:



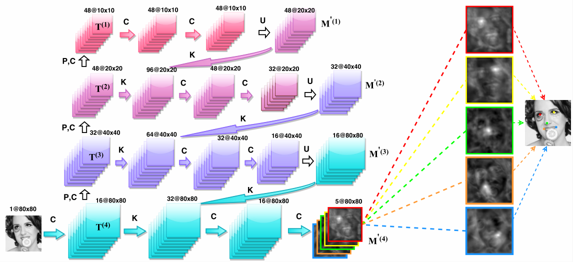
- U is an upsampling layer
- K is concatenation along feature maps dimension
- C is a convolutional layer
- P is a pooling layer

SumNet vs. RCN

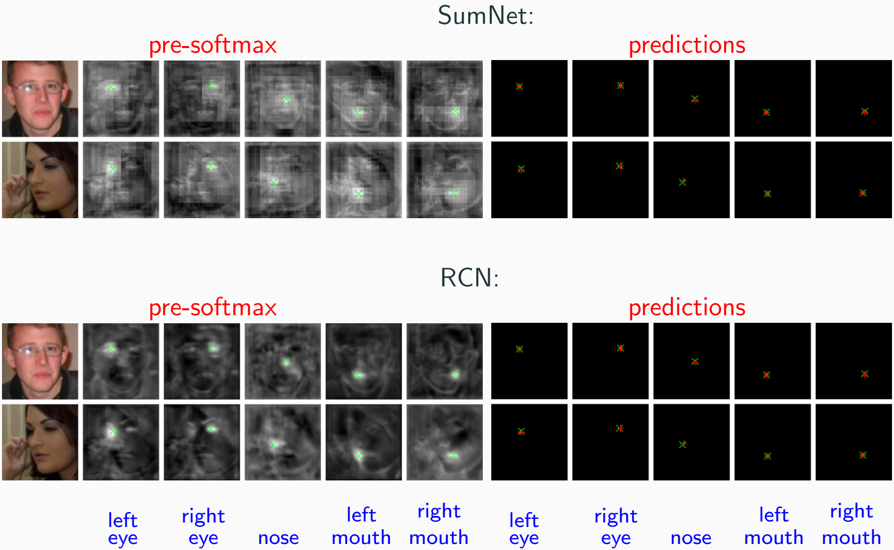
Summation-based Networks (SumNet)



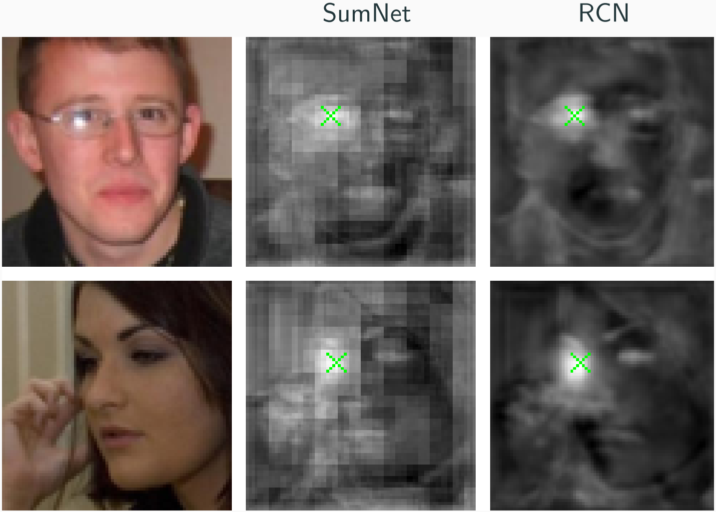
Recombinator Networks (RCN)



SumNet vs. RCN Maps



SumNet vs. RCN Pre-Softmax Maps



Evaluation Datasets (5 keypoints)

Training set:

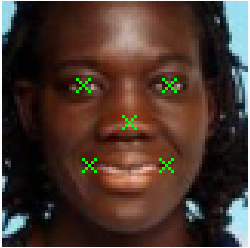
- 10,000 training images
- Data-augmentation: random scale, translation and rotation

Test set:

- AFLW (2995 images)
- AFW (377 images)

For each image 5 keypoints are given:

left eye, right eye, nose, left mouth, right mouth



Comparing SumNet and RCN

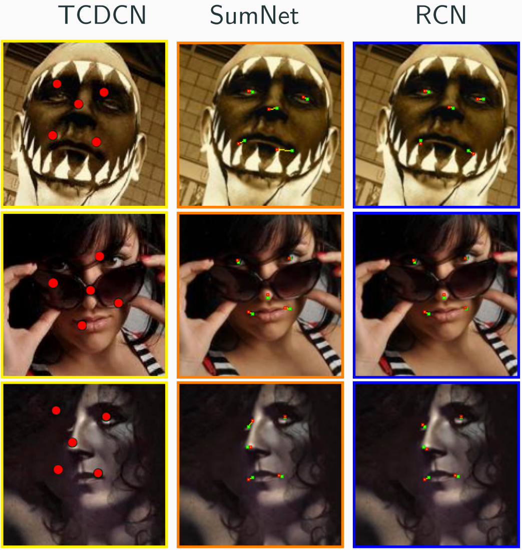
Performance:

Model	AFLW	AFW
SumNet (6 branch - occlusion)	6.27	6.33
RCN (6 branch - occlusion)	**5.60**	**5.36**

Training Time:

- **Convergence:**
  RCN: 200 epochs (4 hours on K20 gpu)
  SumNet: 800 epochs (14 hours on K20 gpu).
- **Reaching error below 7:**
  RCN: 15 epochs (1,050 updates)
  SumNet: 110 epochs (7,800 updates)

Prediction Comparison



Green dots: True key-points, Red dots: Model predictions

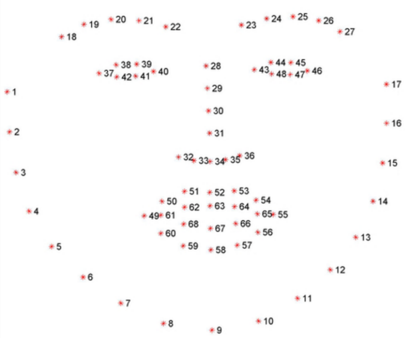
Comparison with Other Models

Model	AFLW	AFW
TSPM [17]	15.9	14.3
CDM [12]	13.1	11.1
ESR [3]	12.4	10.4
RCPR [2]	11.6	9.3
SDM [11]	8.5	8.8
TCDCN [14]	8.0	8.2
TCDCN baseline (our implementation)	7.60	7.87
SumNet (FCN/HC) baseline (this)	6.27	6.33
RCN (this)	**5.60**	**5.36**

Table: Facial landmark estimation error (as a percent; lower is better).

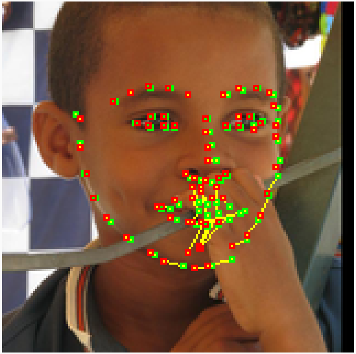
300W Dataset

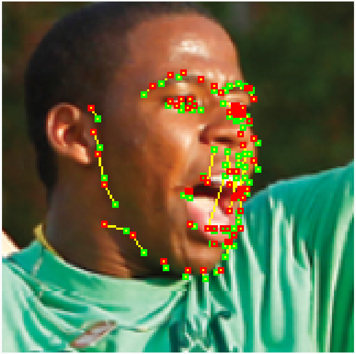
The dataset annotates 68 facial keypoints



- **Train set**: 3148 images (2000 Helen, 811 LFPW, 337 AFW)
- **Test set**: 689 images (330 Helen, 224 LFPW, 135 IBUG)
  - **common subset**: Union of Helen and LFPW test sets
  - IBUG test set contains more extreme pose, expression, and rotation

- Convnet outputs do not always correspond to a plausible keypoint distribution.

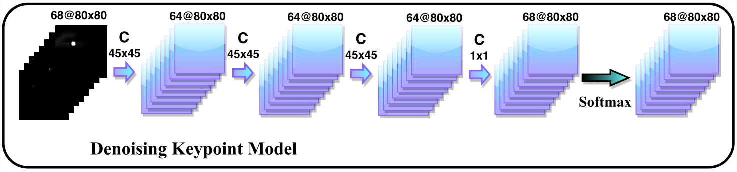




Green dots: True key-points, Red dots: Model predictions, yellow line: connects model prediction to true keypoint.

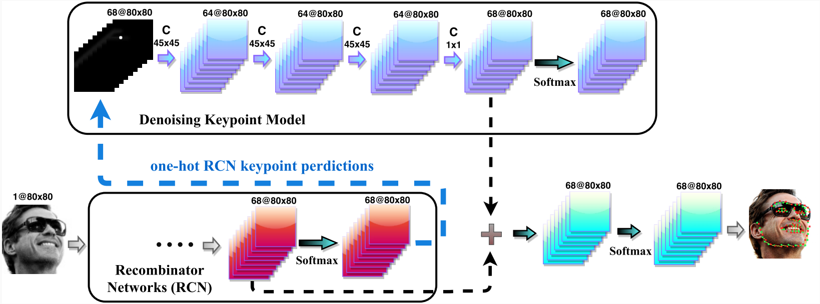
Denoising Keypoint Model

- Each keypoint location is given in a one-hot 2D map
- A subset of keypoint locations are jittered uniformly on the 2D maps
- The model is asked to reconstruct the jittered keypoints



Joint Model

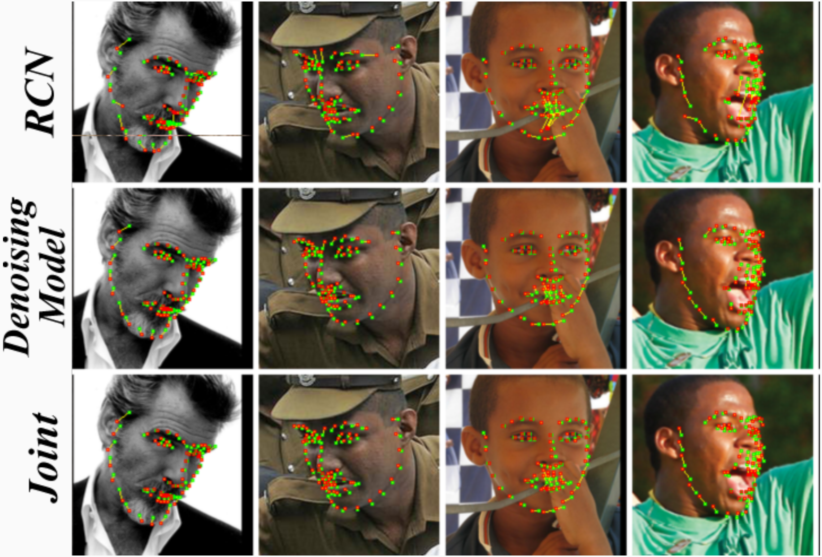
- The Recombinator Networks (RCN) and denoising models are trained separately.
- For prediction:
  1. The keypoint hard prediction of RCN is injected into the denoising model.
  2. The pre-softmax values of RCN and denoising models are summed and pass through a final softmax.



Prediction Samples

Green dots: True key-points, Red dots: Model predictions

Prediction Samples



Green dots: True key-points, Red dots: Model predictions

Prediction Samples

Green dots: True key-points, Red dots: Model predictions

Comparison with Other Models

Model	Common	IBUG	Fullset
CDM [12]	10.10	19.54	11.94
DRMF [1]	6.65	19.79	9.22
RCPR [2]	6.18	17.26	8.35
GN-DPM [10]	5.78	-	-
CFAN [13]	5.50	16.78	7.69
ESR [3]	5.28	17.00	7.58
SDM [11]	5.57	15.40	7.50
ERT [4]	-	-	6.40
LBF [7]	4.95	11.98	6.32
CFSS[16]	4.73	9.98	5.76
TCDCN* [15]	4.80	8.60	5.54
RCN (this)	4.70	9.00	5.54
RCN + denoising model (this)	**4.67**	**8.44**	**5.41**

Table: Facial landmark estimation error (as a percent; lower is better). (* Trained on extra data)

Conclusion

- We propose a model for merging coarse-to-fine features
- The features are injected to finer layers early in their computation
- It improves performance and convergence time
- We propose a convnet-based denoising model for keypoints
- We report SOTA on two 5-keypoint sets and one 68-keypoint set

Questions?

References I

A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic.
Robust discriminative response map fitting with constrained local models.
In *CVPR*, pages 3444–3451, 2013.

X. Burgos-Artizzu, P. Perona, and P. Dollár.
Robust face landmark estimation under occlusion.
In *ICCV*, pages 1513–1520, 2013.

X. Cao, Y. Wei, F. Wen, and J. Sun.
Face alignment by explicit shape regression.
In *ICCV*, 107(2):177–190, 2014.

X. Cao, Y. Wei, F. Wen, and J. Sun.
Face alignment by explicit shape regression.
In *IJCV*, 107(2):177–190, 2014.

References II

B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik.
Hypercolumns for object segmentation and fine-grained localization.
In *CVPR*, 2015.

J. Long, E. Shelhamer, and T. Darrell.
Fully convolutional networks for semantic segmentation.
In *CVPR*, 2015.

S. Ren, X. Cao, Y. Wei, and J. Sun.
Face alignment at 3000 fps via regressing local binary features.
In *CVPR*, pages 1685–1692, 2014.

Y. Sun, X. Wang, and X. Tang.
Deep convolutional network cascade for facial point detection.
In *CVPR*, pages 3476–3483, 2013.

J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler.
Efficient object localization using convolutional networks.
In *CVPR*, 2015.

G. Tzimiropoulos and M. Pantic.
Gauss-newton deformable part models for face alignment in-the-wild.
In *CVPR*, pages 1851–1858, 2014.

X. Xiong and F. De la Torre.
Supervised descent method and its applications to face alignment.
In *CVPR*, pages 532–539, 2013.

References IV

X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas.
Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model.
In *ICCV*, pages 1944–1951, 2013.

J. Zhang, S. Shan, M. Kan, and X. Chen.
Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment.
In *ECCV*, pages 1–16. 2014.

Z. Zhang, P. Luo, C. Loy, and X. Tang.
Facial landmark detection by deep multi-task learning.
In *ECCV*, pages 94–108. 2014.

References V

Z. Zhang, P. Luo, C. C. Loy, and X. Tang.
Learning deep representation for face alignment with auxiliary attributes.
In *PAMI*, 2015.

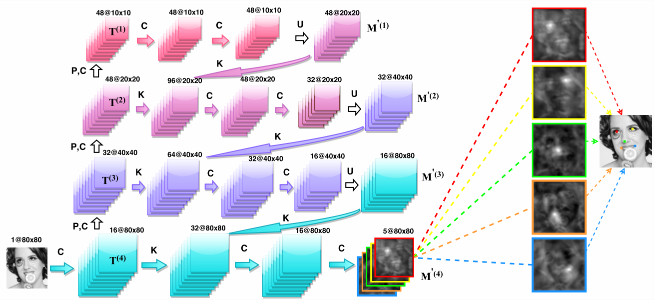
S. Zhu, C. Li, C. C. Loy, and X. Tang.
Face alignment by coarse-to-fine shape searching.
In *CVPR*, pages 4998–5006, 2015.

X. Zhu and D. Ramanan.
Face detection, pose estimation, and landmark localization in the wild.
In *CVPR*, pages 2879–2886, 2012.

Training Cost

$$\mathcal{L}(\mathbf{W}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} -\log P(z_k^{(n)}|u^{(n)}) + \lambda\|\mathbf{W}\|^2 \qquad (1)$$

- $u^{(n)}$: input image n
- $z_k^{(n)}$: target location for keypoint k in image n
- $\mathbf{W}$: network parameters



Error Metric

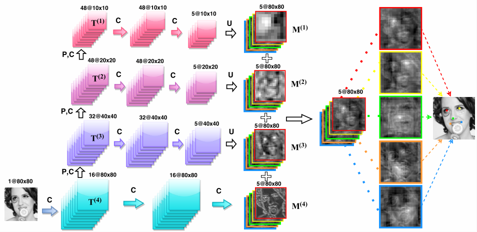
Euclidean distance between the true and estimated landmark positions normalized by the distance between the eyes (interocular distance):

$$\text{error} = \frac{1}{KN} \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\sqrt{(x_k^{(n)} - \tilde{x}_k^{(n)})^2 + (y_k^{(n)} - \tilde{y}_k^{(n)})^2}}{D^{(n)}} \tag{2}$$

- $(x_k^{(n)}, y_k^{(n)})$: true x and y coordinates for keypoint k in image n
- $(\tilde{x}_k^{(n)}, \tilde{y}_k^{(n)})$: model predicted coordinates
- $D^{(n)}$: interocular distance in image n

Masking SumNet Branches

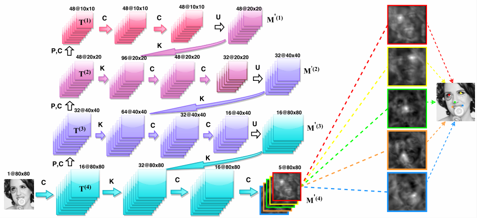
Mask	AFLW	AFW
coarse → fine		
1, 0, 0, 0	10.54	10.63
0, 1, 0, 0	11.28	11.43
1, 1, 0, 0	9.47	9.65
0, 0, 1, 0	16.14	16.35
0, 0, 0, 1	45.39	47.97
0, 0, 1, 1	13.90	14.14
0, 1, 1, 1	7.91	8.22
1, 0, 0, 1	6.91	7.51
1, 1, 1, 1	**6.44**	**6.78**



Mask: 0 branch is omitted, 1 branch in included.

- Error values are in percent

Mask	AFLW	AFW
coarse → fine		
1, 0, 0, 0	10.61	10.89
0, 1, 0, 0	11.56	11.87
1, 1, 0, 0	9.31	9.44
0, 0, 1, 0	15.78	15.91
0, 0, 0, 1	46.87	48.61
0, 0, 1, 1	12.67	13.53
0, 1, 1, 1	7.62	7.95
1, 0, 0, 1	6.79	7.27
1, 1, 1, 1	**6.37**	**6.43**

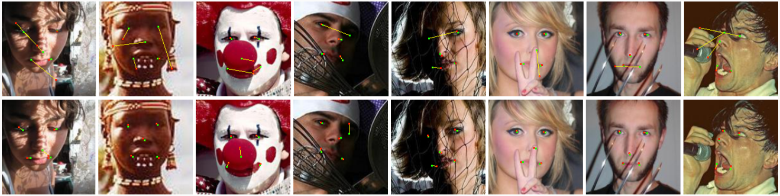


Mask: 0 branch is omitted. 1 branch in included.

- Error values are in percent

Occlusion Pre-Processing

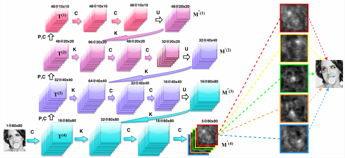
- Test sets contain more extreme occlusion and lighting cotrast
- We put black rectangle on random location in the image



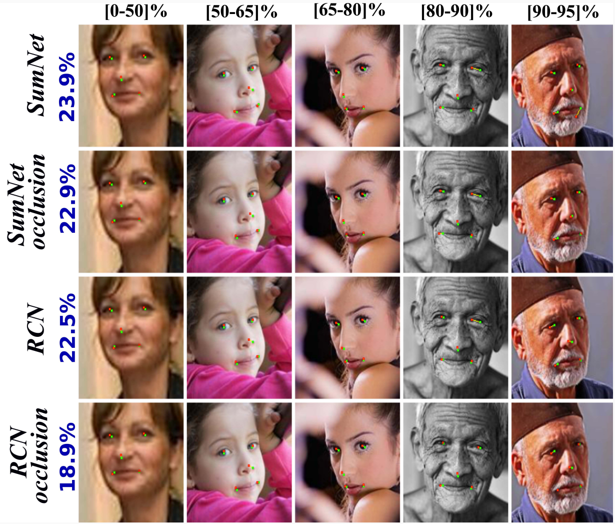
This forces the model to look at more global facial components

Adding More Branches

Model	AFLW	AFW
SumNet (4 branch)	6.44	6.78
SumNet (5 branch)	6.42	6.53
SumNet (6 branch)	6.34	6.48
SumNet (5 branch - occlusion)	6.29	6.34
SumNet (6 branch - occlusion)	**6.27**	**6.33**
RCN (4 branch)	6.37	6.43
RCN (5 branch)	6.11	6.05
RCN (6 branch)	6.00	5.98
RCN (7 branch)	6.17	6.12
RCN (5 branch - occlusion)	5.65	5.44
RCN (6 branch - occlusion)	**5.60**	**5.36**
RCN (7 branch - occlusion)	5.76	5.55

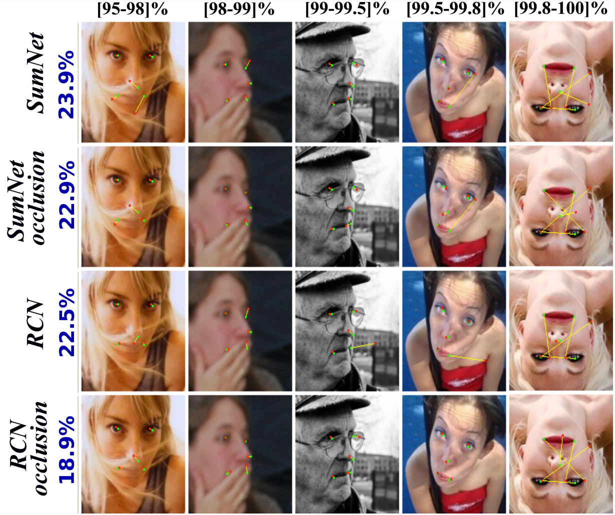


Prediction Results, part 1



- For each image in test sets average error is taken (across 4 models)
- The images are sorted (by avg error) and a random sample is taken in each bin

Prediction Results, part 2



- For each image in test sets average error is taken (across 4 models)
- The images are sorted (by avg error) and a random sample is taken in each bin

Comparison with Other Architectures

Features \ Models	Efficient Localization [9]	Deep Cascade [8]	Hyper-columns [5]	FCN [6]	RCN (this)
Coarse features: hard crop or soft combination?	Hard	Hard	Soft	Soft	Soft
Learned coarse features fed into finer branches?	No	No	No	No	Yes

Table: Comparison of multi-resolution architectures. The Efficient Localization and Deep Cascade models use coarse features to crop images (or fine layer features), which are then fed into fine models. This process saves computation when dealing with high-resolution images but at the expense of making a greedy decision halfway through the model. Soft models merge local and global features of the entire image and do not require a greedy decision. The Hypercolumn and FCN models propagate all coarse information to the final layer but merge information via addition instead of conditioning fine features on coarse features. The Recombinator Networks (RCN), in contrast, injects coarse features directly into finer branches, allowing the fine computation to be tuned by (conditioned on) the coarse information.