Montreal Institute for Learning Algorithms MILA

Recombinator Networks: Learning Coarse-to-Fine Feature Aggregation

Sina Honari [1], Jason Yosinski [2], Pascal Vincent [1], Christopher Pal [3]

[1] University of Montreal, [2] Cornell University, [3] Ecole Polytechnique of Montreal

Université de Montréal

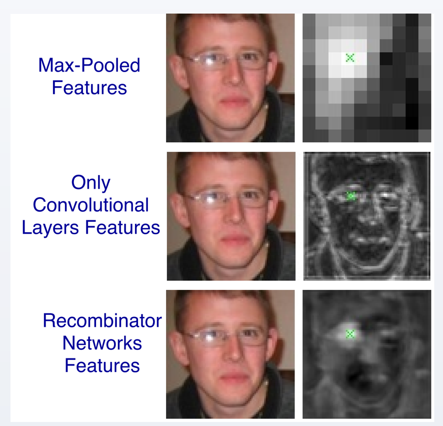
Motivation

Convnets are usually composed of alternating convolutional and max-pooling layers:

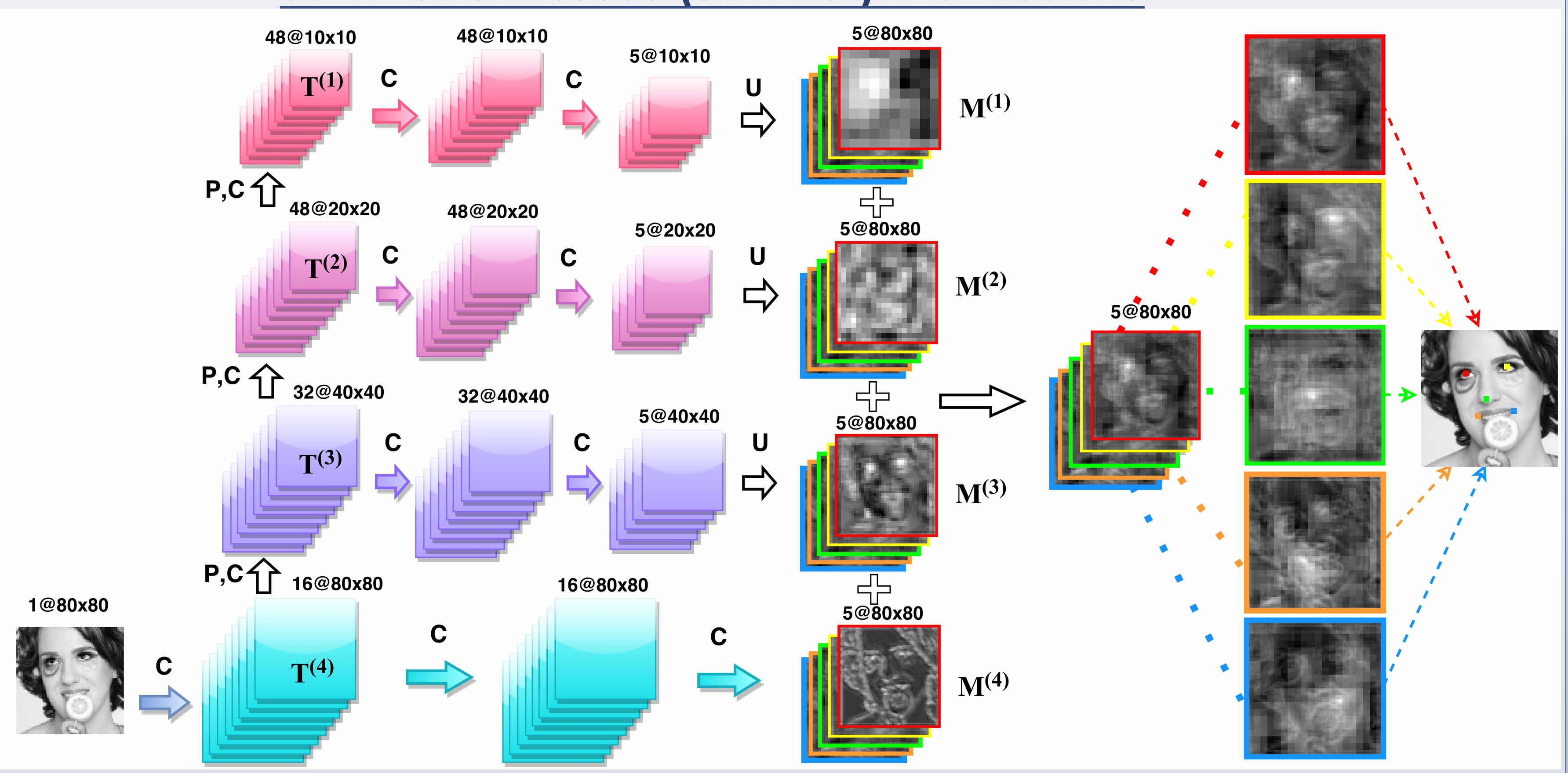
Max-pooling: Gets robust features, but loses precise spatial information

Network of only convolutional layers: keeps spatial information, but has lots of false positives (regions of high probability).

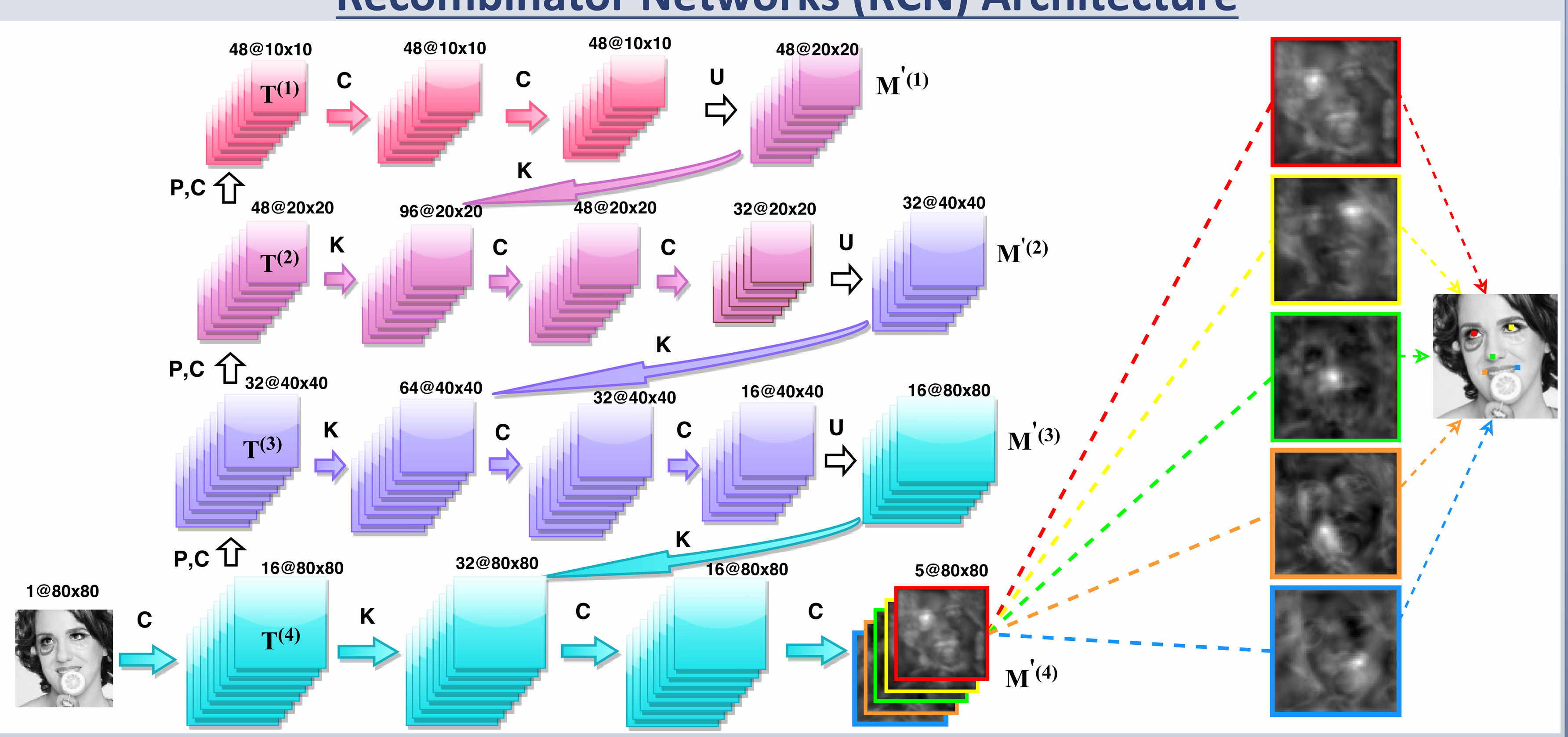
Is there a way to use robust pooled features and meanwhile keep positional information?



Summation–based (SumNet) Architecture



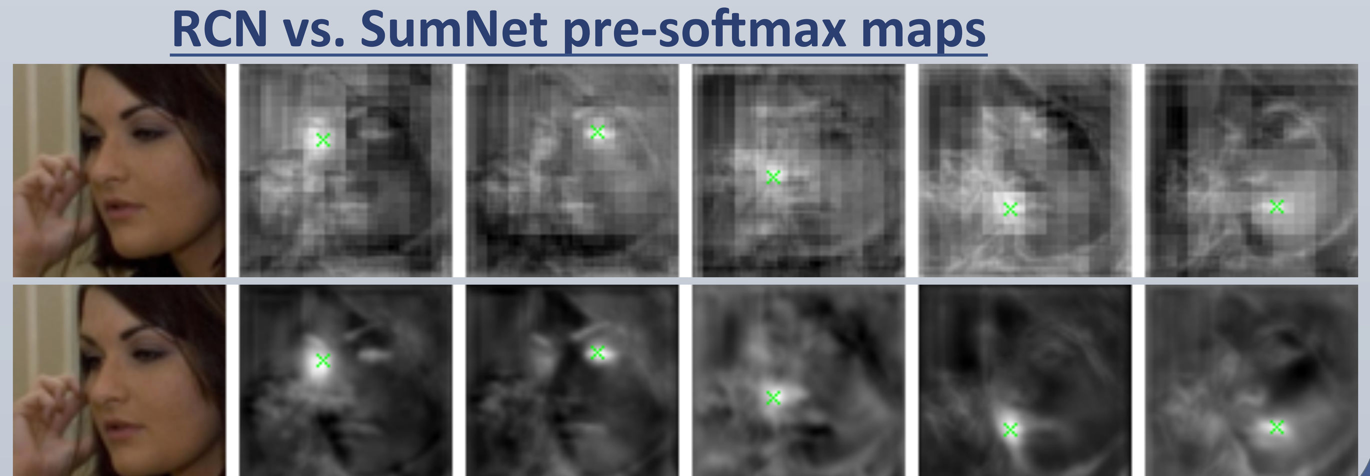
Recombinator Networks (RCN) Architecture



RCN vs. SumNet pre-softmax maps

SumNet Maps

RCN Maps



Experimental Results

Cost:

$$\mathcal{L}(\mathbf{W}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} -\log P(Y_k = y_k^{(n)} | X = x^{(n)}) + \lambda\|\mathbf{W}\|^2.$$

Error:

$$\frac{1}{KN}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\sqrt{(w_k^{(n)} - \tilde{w}_k^{(n)})^2 + (h_k^{(n)} - \tilde{h}_k^{(n)})^2}}{D^{(n)}},$$

	SumNet		RCN	
Mask	AFLW	AFW	AFLW	AFW
1, 0, 0, 0	10.54	10.63	10.61	10.89
0, 1, 0, 0	11.28	11.43	11.56	11.87
1, 1, 0, 0	9.47	9.65	9.31	9.44
0, 0, 1, 0	16.14	16.35	15.78	15.91
0, 0, 0, 1	45.39	47.97	46.87	48.61
0, 0, 1, 1	13.90	14.14	12.67	13.53
0, 1, 1, 1	7.91	8.22	7.62	7.95
1, 0, 0, 1	6.91	7.51	6.79	7.27
1, 1, 1, 1	**6.44**	**6.78**	**6.37**	**6.43**

Table 1. The performance of SumNet and RCN trained with masks applied to different branches. A mask value of 1 indicates the branch is included in the model and 0 indicates it is omitted (as a percent; lower is better). In SumNet model mask 0 indicates no contribution from that branch to the summation of all branches, while in RCN, if a branch is omitted, the previous coarse branch is upsampled to the following fine branch. The mask numbers are ordered from the coarsest branch to the finest branch.

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
RCN (6 branch - occlusion - skip)	5.63	5.56

Table 2. SumNet and RCN performance with different number of branches, occlusion preprocessing and skip connections.

5 keypoint Datasets:

Training Set: (MTFL Dataset; 9000 train – 1000 valid), Testing Set: 2995 AFLW, 337 AFW

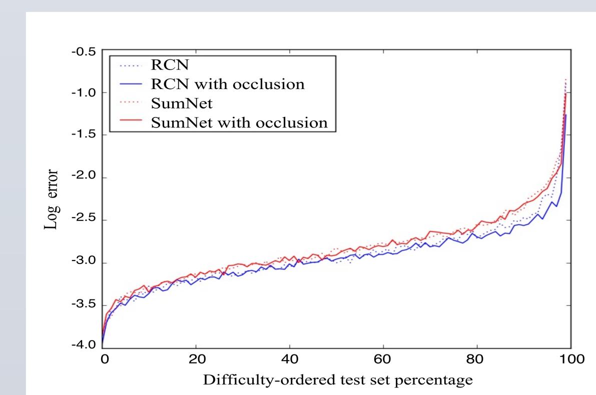
Model	AFLW	AFW
TSPM [45]	15.9	14.3
CDM [38]	13.1	11.1
ESR [6]	12.4	10.4
RCPR [5]	11.6	9.3
SDM [35]	8.5	8.8
TCDCN [41]	8.0	8.2
TCDCN baseline (our implementation)	7.60	7.87
SumNet (FCN/HC) baseline (this)	6.27	6.33
RCN (this)	**5.60**	**5.36**

Table 3. Facial landmark mean error normalized by interocular distance on AFW and AFLW sets (as a percent; lower is better).[11]

Training Time:

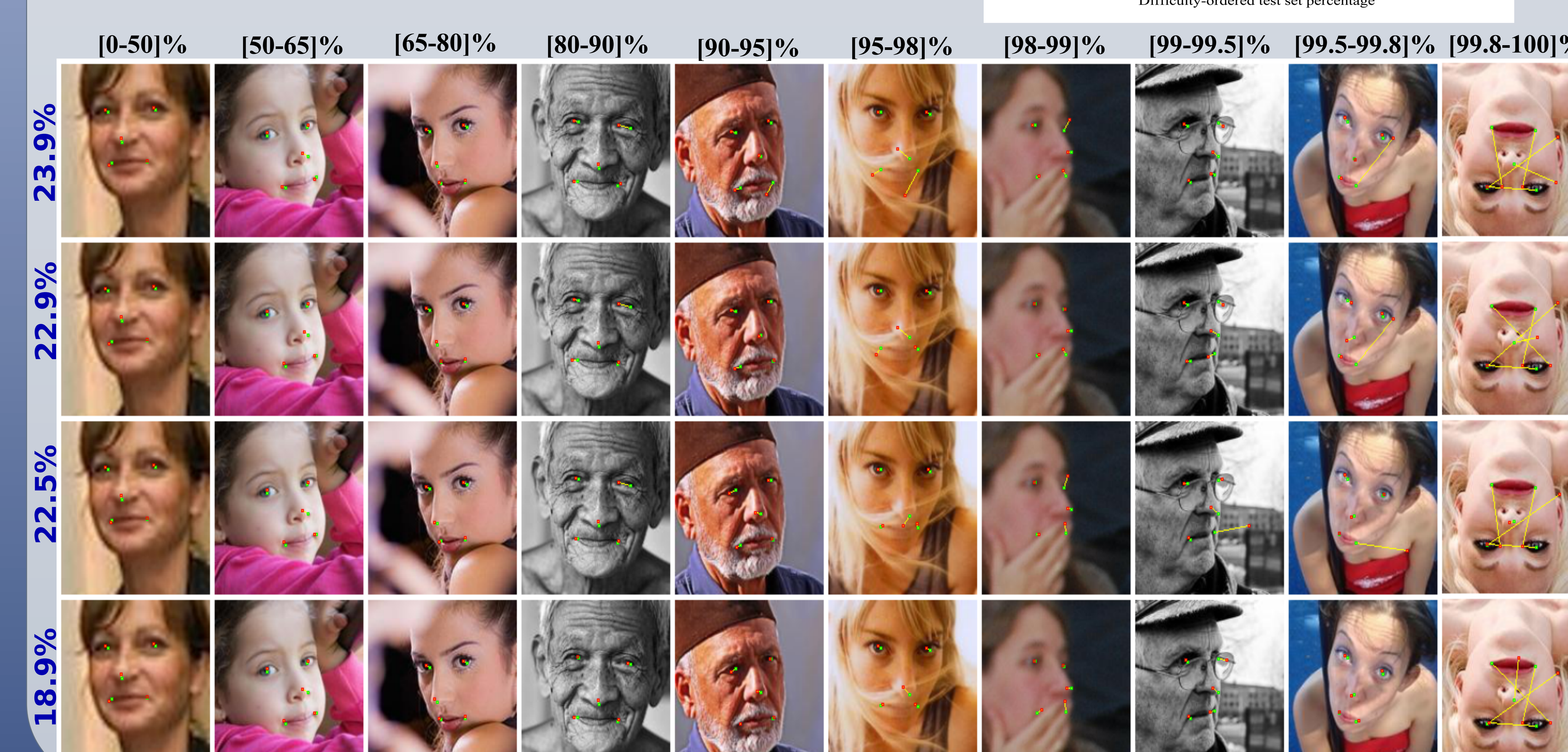
RCN trains faster

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



Occlusion Preprocessing:

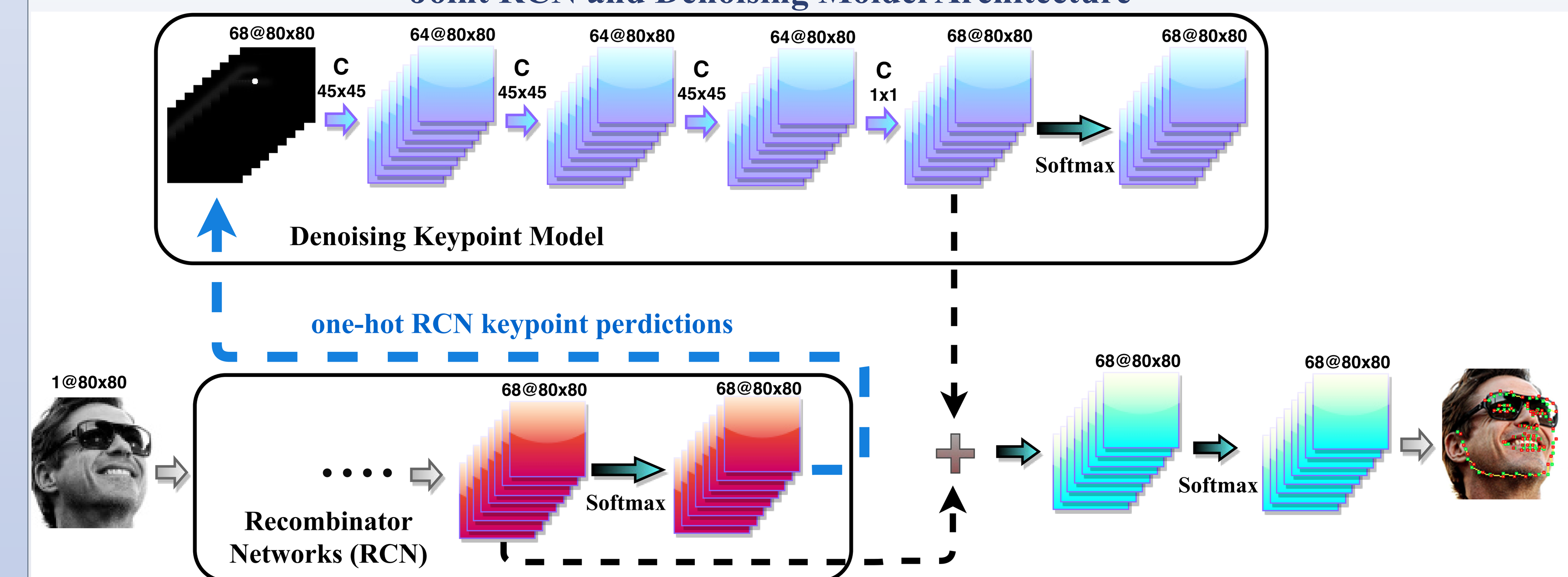
Black patches of variable size (20 to 50 pixels) is put on top of the images in random locations to force the model to look at more global features.



300W Dataset:

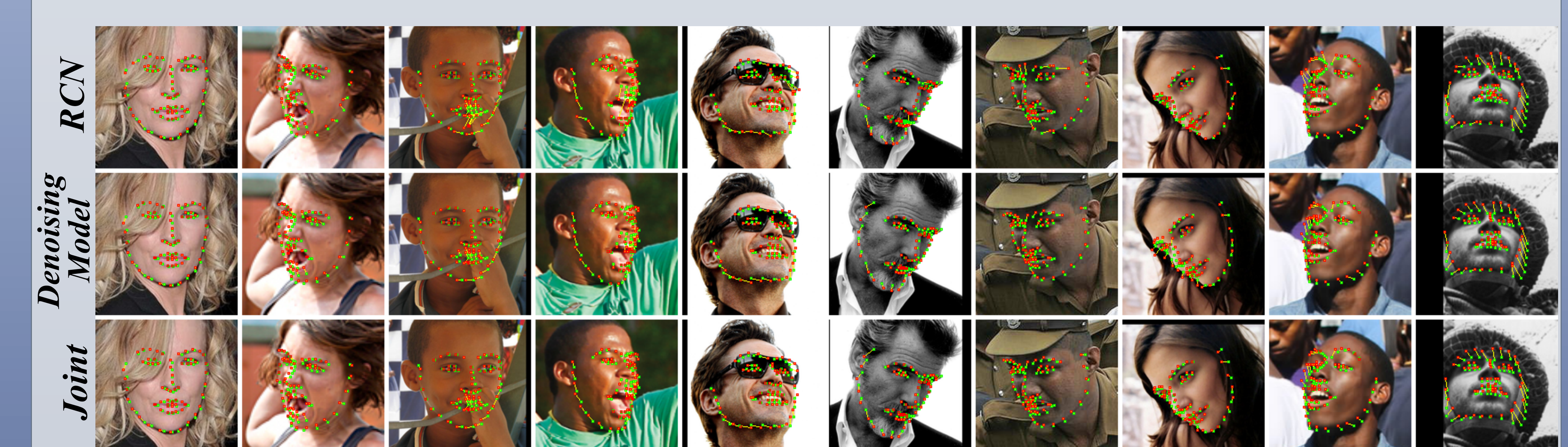
Train Set: (337 AFW , 2000 Helen, 811 LFPW), Test Set: 135 iBug, 224 LFPW, 330 Helen

Joint RCN and Denoising Moldel Architecture



Model	#keypoints	Common	IBUG	Fullset
PO-CR [32]		4.00	6.82	4.56
RCN (this)	49	2.64	5.10	3.88
RCN + denoising keypoint model (this)		**2.59**	**4.81**	**3.76**
CDM [38]		10.10	19.54	11.94
DRMF [2]		6.65	19.79	9.22
RCPR [5]		6.18	17.26	8.35
GN-DPM [33]		5.78	-	-
CFAN [40]		5.50	16.78	7.69
ESR [6]		5.28	17.00	7.58
SDM [35]	68	5.57	15.40	7.50
ERT [7]		-	-	6.40
LBF [18]		4.95	11.98	6.32
CFSS[44]		4.73	9.98	5.76
TCDCN † [42]		4.80	8.60	5.54
RCN (this)		4.70	9.00	5.54
RCN + denoising keypoint model (this)		**4.67**	**8.44**	**5.41**

Table 4. Facial landmark mean error normalized by interocular distance on 300W test sets (as a percent; lower is better).[11]



Features \ Models	Efficient Localization [31]	Deep Cascade [28]	Hyper-columns [13]	FCN [17]	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 5. 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. The model is trained end-to-end and results in *learned* coarse features which are tuned directly to support the eventual fine predictions.