Improving Face Recognition by Exploring Local Features with Visual Attention

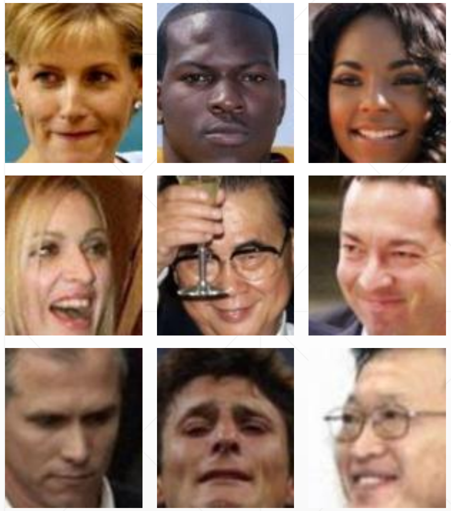
Yichun Shi and Anil K. Jain

Michigan State University

Difficulties of Face Recognition

- Large variations in unconstrained face datasets
- Face alignment partially solve the problem
- Variations still remains after alignment

Example face images after alignment



LFW

IJB-B

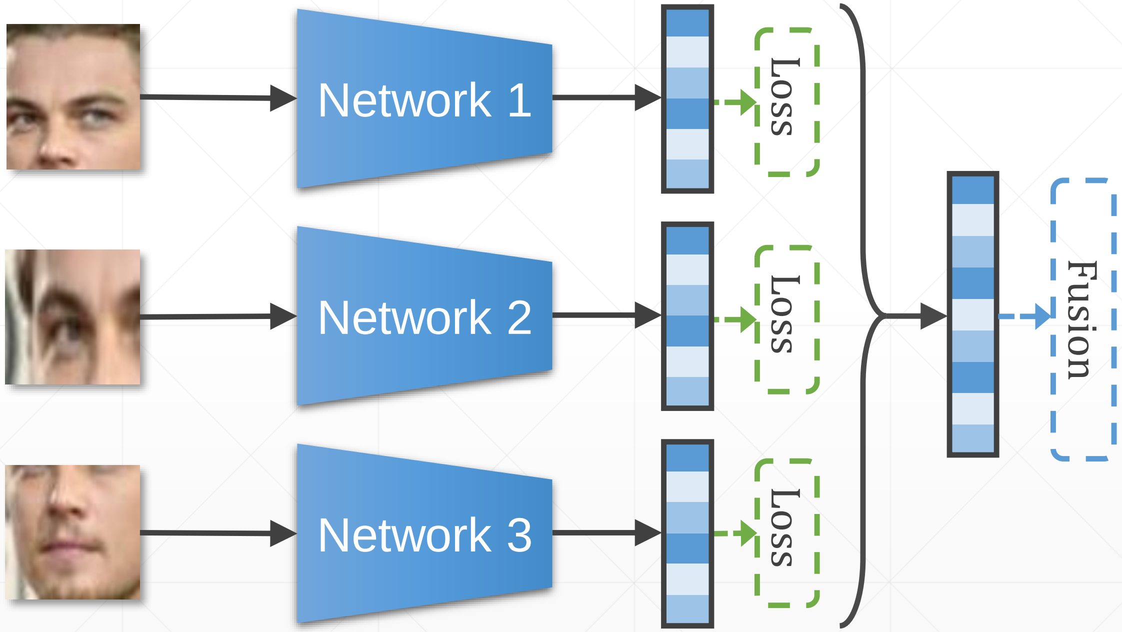
Parts-based Method

- Cropping patches for different facial parts [1]
- Building models for different patches
- Fuse the representations or scores

➢ **Problems**:

- Deciding useful facial parts
- Learning complementary features
- Effective fusion

➢ **An end-to-end solution**



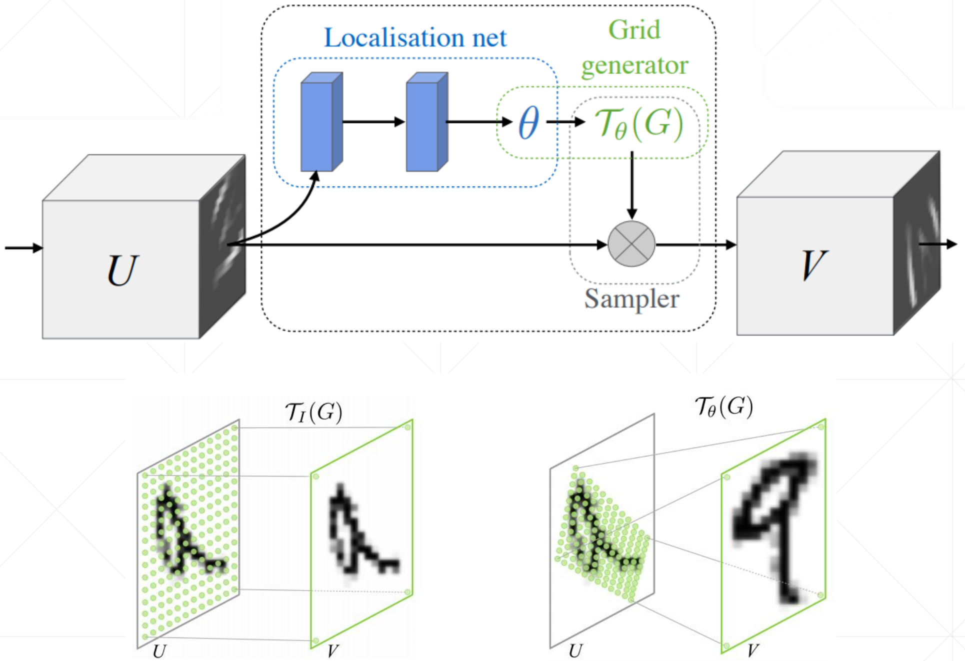
[1] Y. Sun, X. Wang, and X. Tang. “Deep learning face representation from predicting 10,000 classes”. In CVPR, 2014.

Spatial Transformer Network

- An *Attention Network* predicts transformation matrix θ
- A grid sampler transforms the image:

$$\begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} = \frac{1}{\lambda}\begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & 1 \end{pmatrix}\begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$

- Bilinear sampling
- **Differentiable**

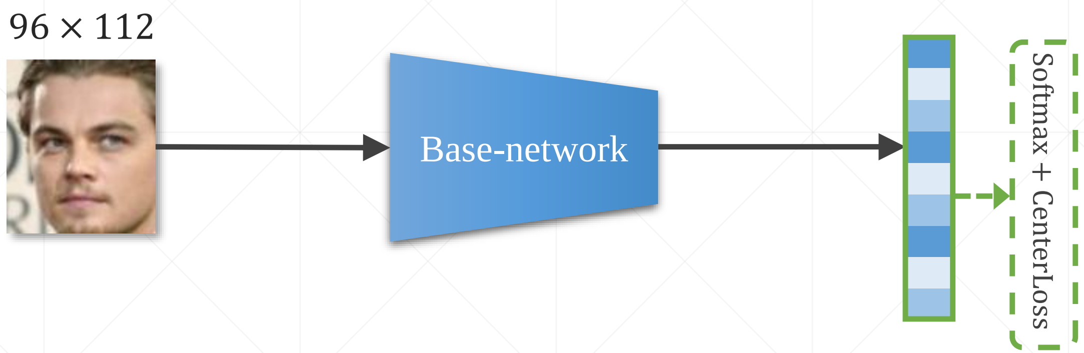


Spatial Transformer [1]

[1] M. Jaderberg, K. Simonyan, A. Zisserman, et al. “Spatial Transformer Networks.” In NIPS, 2015.

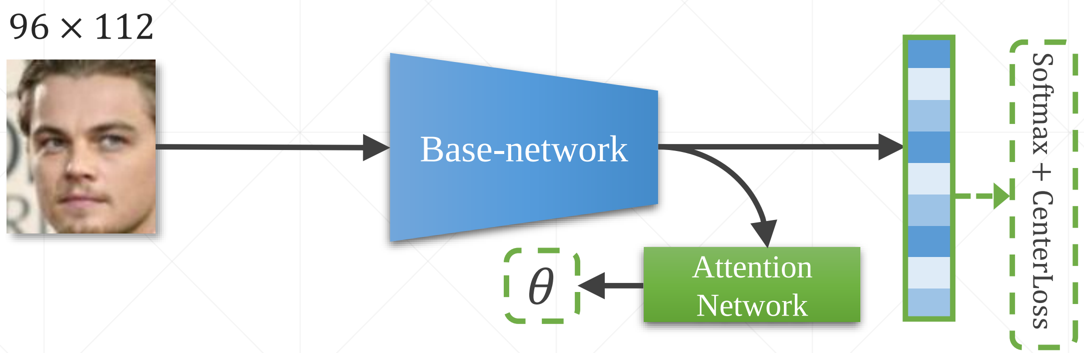
Architecture – Baseline

- Build upon a typical single CNN system
- Pre-aligned input (112x96)
- Base-net: any CNN



Architecture – Attention Network

- Spatial Transformer Network
- Last feature map as input
- Predicts K transformation matrices θ



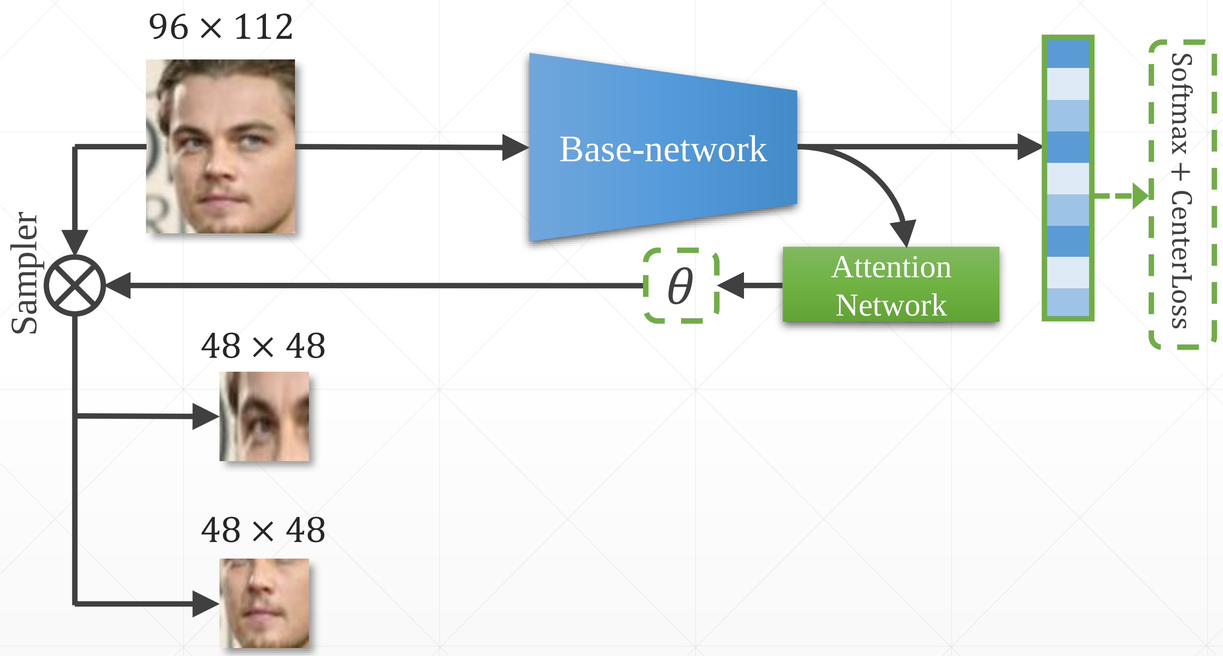
Type	Output Size
Batch Norm + Fully Connected	128
Batch Norm + Fully Connected	8 × K

Architecture of Attention Network

Architecture – Attention Network

- Grid sampler using θ
- K output patches

  ($K = 2$ in example figure)
- 48×48 size

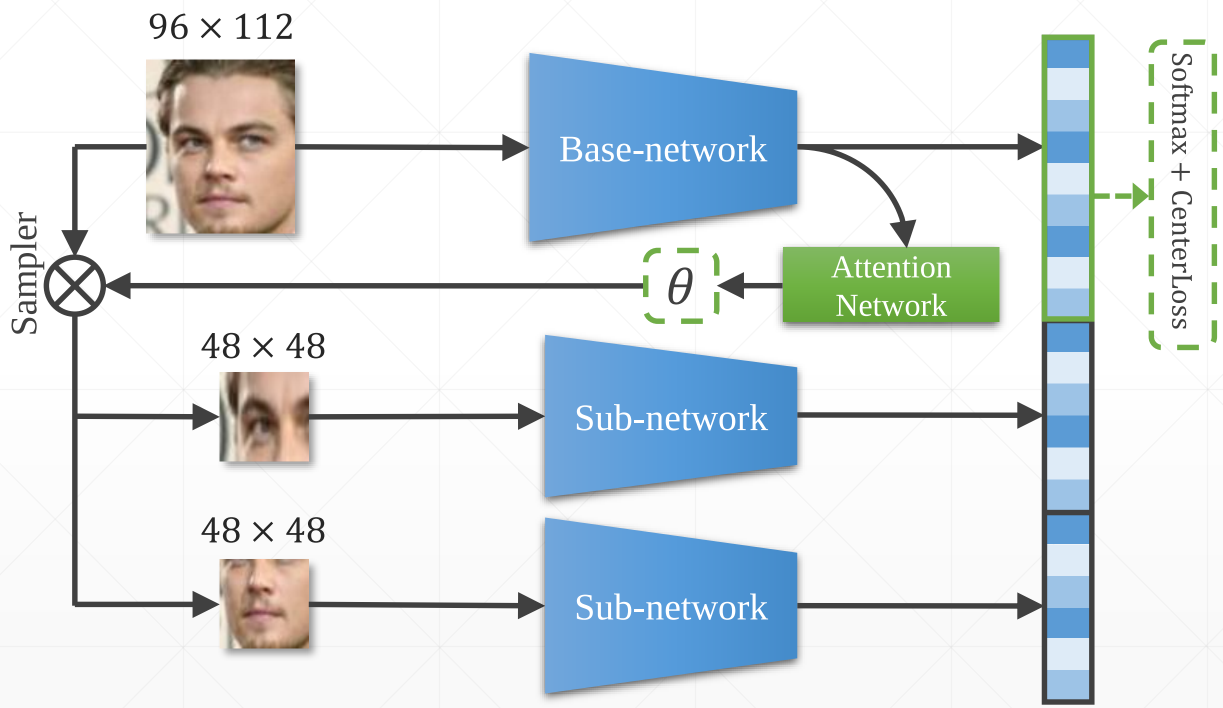


Architecture – Sub-network

- K subnetworks
- Each learns a local feature vector

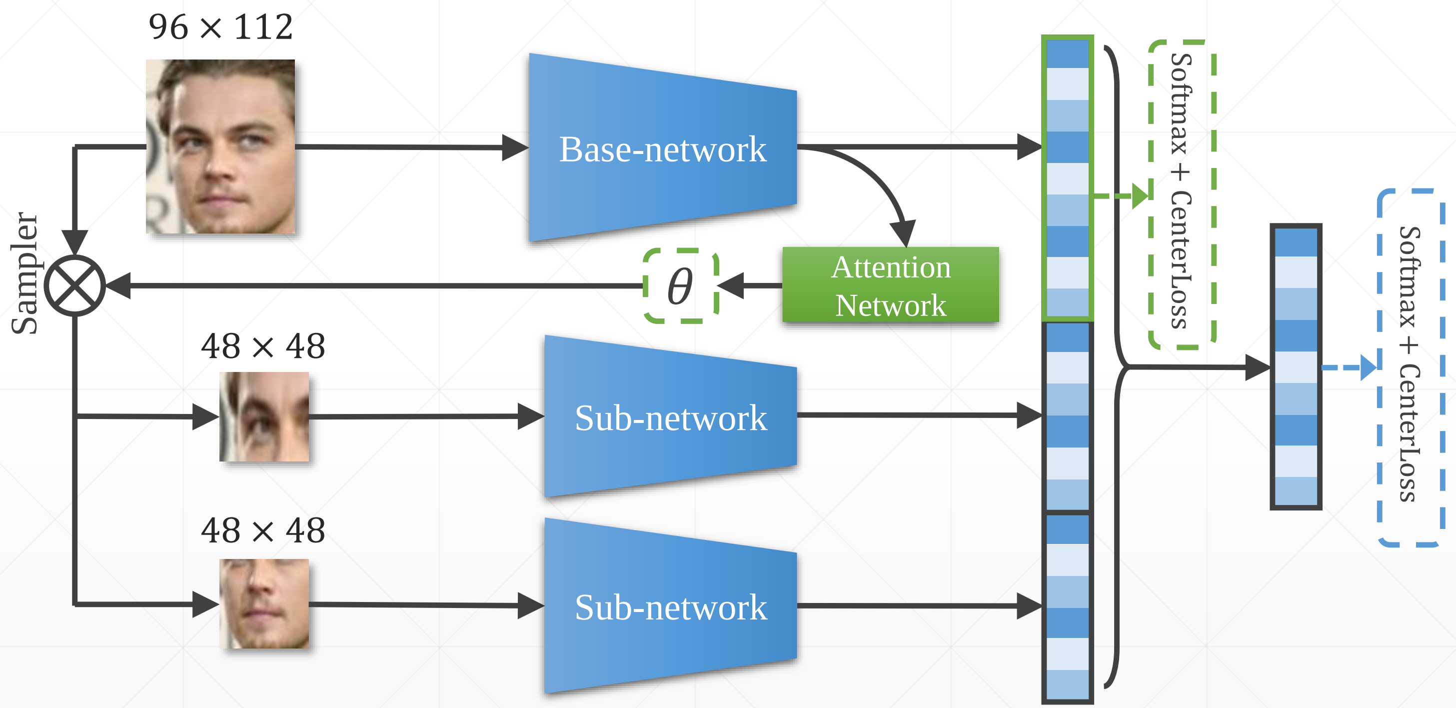
Type	Output Size	Filter Size/Stride
Convolution	48×48×32	3×3/1
Convolution	48×48×64	3×3/1
Max Pooling	24×24×64	2×2/2
Convolution	24×24×64	3×3/1
Convolution	24×24×128	3×3/1
Max Pooling	12×12×128	2×2/2
Convolution	12×12×128	3×3/1
Convolution	12×12×256	3×3/1
Max Pooling	6×6×256	2×2/2
Convolution	6×6×256	3×3/1
Convolution	6×6×512	3×3/1
Fully Connected	128	2×2/2

Architecture of Sub-network



Architecture – Fusion Layer

- A fully connected layer to fuse the features
- Classification/Verification loss for the fused feature



Promoting Sub-networks

- Some sub-networks has a very small weight in the fusion layer (dead)
- Dead sub-networks fails to learn useful local features
- Fusion layer:

$$\mathbf{y} = W^g \mathbf{x}^g + W^l \mathbf{x}^l + \mathbf{b}$$

W^g: weights for global (base-network) features
$\mathbf{x}^g$: global features
W^l: weights for local (sub-network) features
$\mathbf{x}^g$: local features
$\mathbf{b}$: biases

- Promotion Loss [1]:

$$\mathcal{L}_p = \frac{1}{D_l} \sum_{i=1}^{D_l} \left\| \left\| W_i^l \right\|^2 - \alpha \right\|^2$$

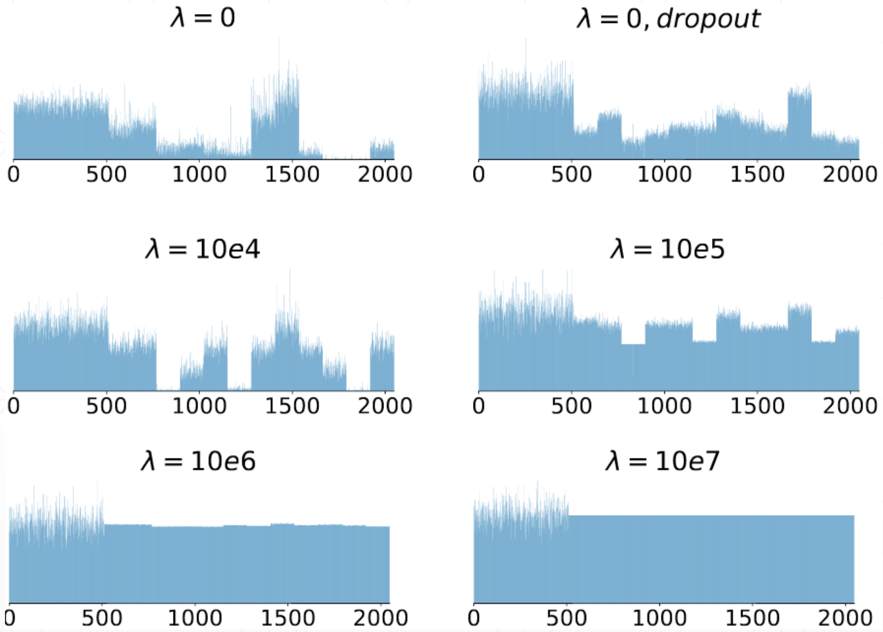
$$\alpha = \frac{1}{D_g} \sum_{i=1}^{D_g} \left\| W_i^l \right\|^2$$

D_l: # dimensions of global feature vector
W_i^l: weight for i_{th} local feature
D_g: # dimensions of global feature vector
W_i^g: weight for i_{th} global feature

[1] Guo and L. Zhang. “One-shot face recognition by promoting underrepresented classes”. arXiv:1707.05574, 2017.

Promoting Sub-networks

- Visualization of magnitude of the weights in fusion layer
- λ: the coefficient of the promotion loss
- Dead neurons without promotion
- Dropout harms the performance



Experiments

- Base-net: Face-ResNet
- Three models:
  - Base-net: $K = 0$, typical single CNN system
  - Model A: $K = 3$, manually initialized patches.
  - Model B: $K = 12$, randomly initialized
- Training Data: CASIA-Webface (0.5M)
- 2/4 GPUs for training Model A/B, respectively
- Inference speed:
  - Base-net: 0.003s per image
  - Model A: 0.003s per image
  - Model B: 0.004s per image

[1] A. Hasnat, J. Bohne, S. Gentric, and L. Chen. "Deepvisage: Making face recognition simple yet with powerful generalization skills". arXiv:1703.08388, 2017.

Results on LFW

Model	AN	FL	PL	Accuracy	VR @ FAR=0.1%	DIR Rank-1 @ FAR=1%
Base-net				98.77%	94.96%	72.96%
Model A	Y	Y	Y	98.85%	95.90%	77.51%
Model B	N	Y	Y	98.67%	95.54%	74.33%
Model B	Y	N	Y	98.78%	95.63%	76.37%
Model B	Y	Y	N	98.75%	95.83%	75.75%
Model B	Y	Y	Y	**98.98%**	**96.44%**	**77.96%**

- AN: Attention Network
- FL: Fusion Layer
- PL: Promotion Loss

- Accuracy: standard LFW protocol
- VR, DIR: BLUFR protocol [1]

[1] S. Liao, Z. Lei, D. Yi, and S. Z. Li. "A benchmark study of large-scale unconstrained face recognition". In IJCB, 2014.

Example localized facial parts

- Consistent localization
- Invariant to variation
- Distinctive regions
- No landmarks are used

Results on IJB-A and IJB-B

Model	TAR@FAR (Verification)		CMC (Closed-set Identification)		FNIR (Open-set Identification)	
	0.001	0.01	Rank-1	Rank-5	0.01	0.1
Base-net	0.542 ± 0.092	0.788 ± 0.092	0.882 ± 0.019	0.954 ± 0.008	0.426 ± 0.017	0.355 ± 0.014
Model A	0.583 ± 0.084	0.808 ± 0.026	0.889 ± 0.007	0.957 ± 0.007	0.418 ± 0.015	**0.353 ± 0.014**
Model B	**0.602 ± 0.069**	**0.823 ± 0.022**	**0.898 ± 0.009**	**0.960 ± 0.006**	**0.411 ± 0.016**	0.353 ± 0.014

Results on IJB-A 1:1 Comparison and 1:N Search protocol

Model	TAR@FAR (Verification)		CMC (Closed-set Identification)		FNIR (Open-set Identification)	
	0.001	0.01	Rank-1	Rank-5	0.01	0.1
Base-net	0.631	0.851	0.749	0.861	0.149	0.032
Model A	0.652	0.861	0.768	**0.875**	0.139	**0.031**
Model B	**0.659**	**0.865**	**0.769**	0.874	**0.135**	0.032

Results on IJB-B 1:1 Baseline Verification and 1:N Mixed Media Identification protocol

Conclusion

- End-to-end Parts-based face recognition
- Automatic localization of facial parts via attention network
- Simultaneous learning of fusion layer

Conclusion

- End-to-end Parts-based face recognition
- Automatic localization of facial parts via attention network
- Simultaneous learning of fusion layer

Future work:

- Architecture of sub-network – efficiency, effectiveness
- More complementary local features

Q/A