End-to-End Localization and Ranking for Relative Attributes

Krishna Kumar Singh and Yong Jae Lee
University of California, Davis

Experiment and presentation by Santiago Gonzalez and Wei-Jen Ko

Agenda

Brief paper review

Code walk through

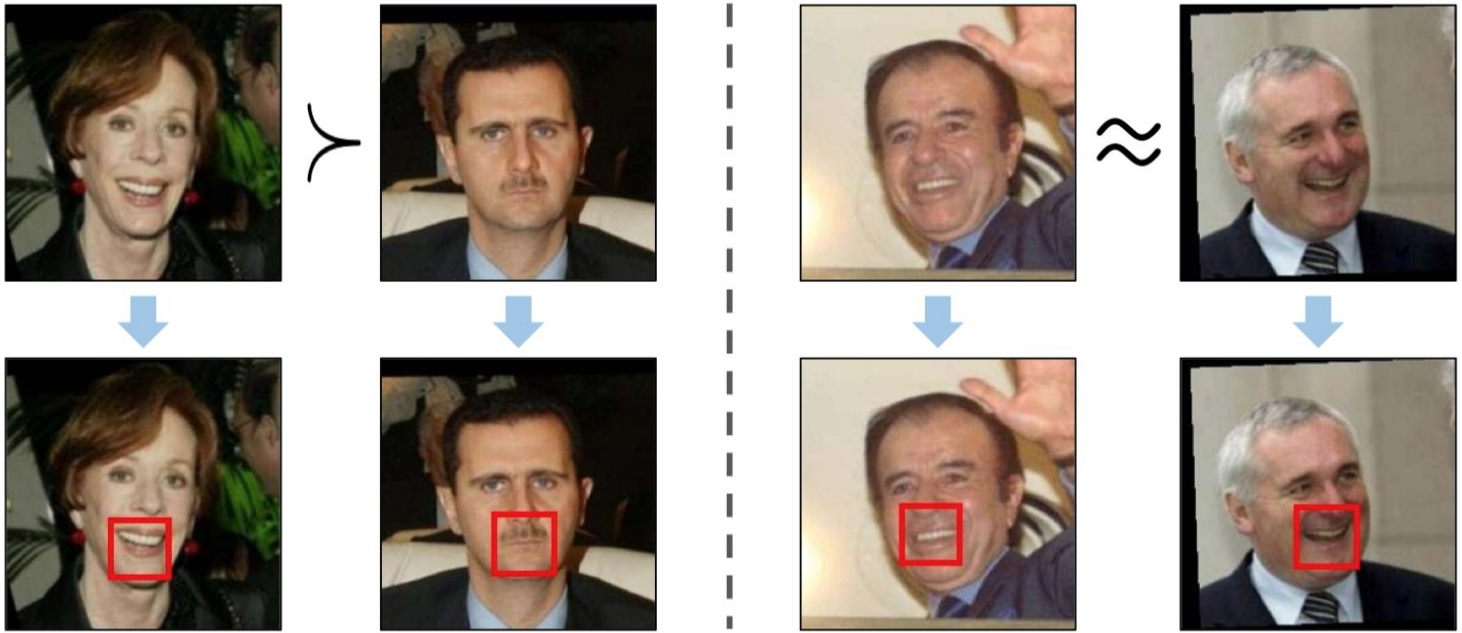
Experiment

Results

Discussion

The Task at Hand

Attribute:
Smile



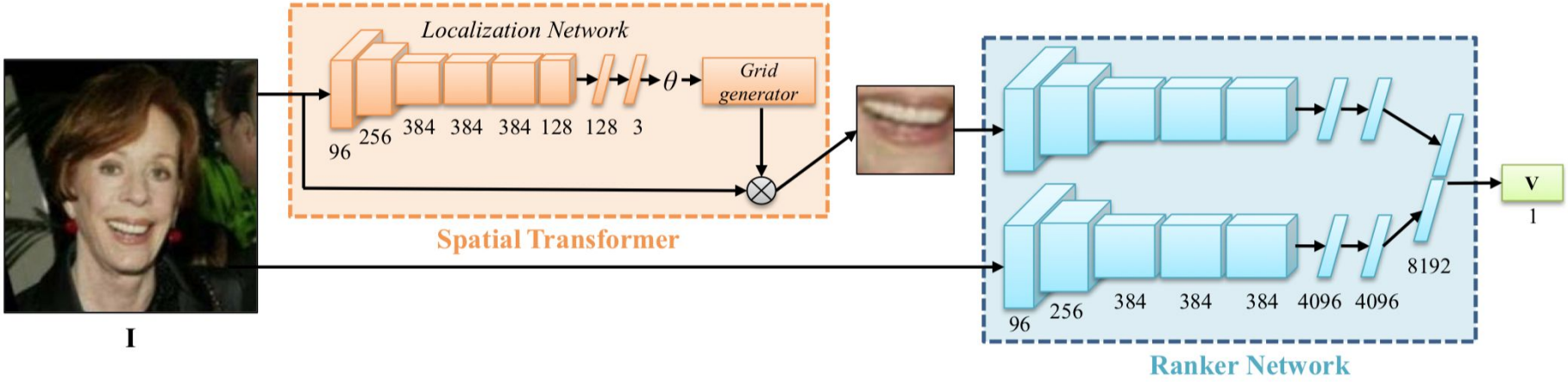
Why is this important?

Similar problem in concept to object affordances: we want to define something as the sum of its parts, rather than as a “unique” entity

Prior recognition systems know objects *a priori*, want to reach zero-shot learning

More fine-grained information than object categories / parts alone

Network Structure

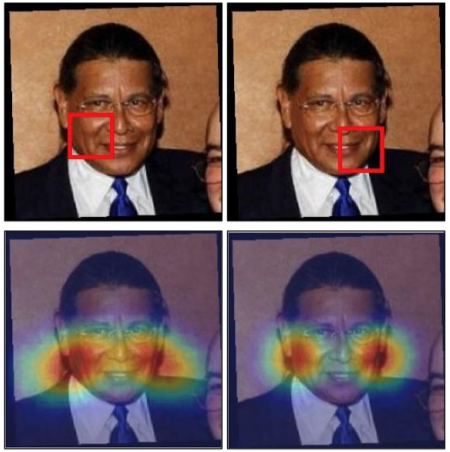


Attribute:
Dark hair

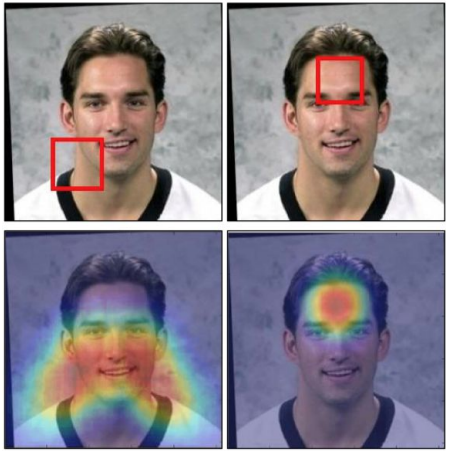
Attribute:
Smile

Training epochs

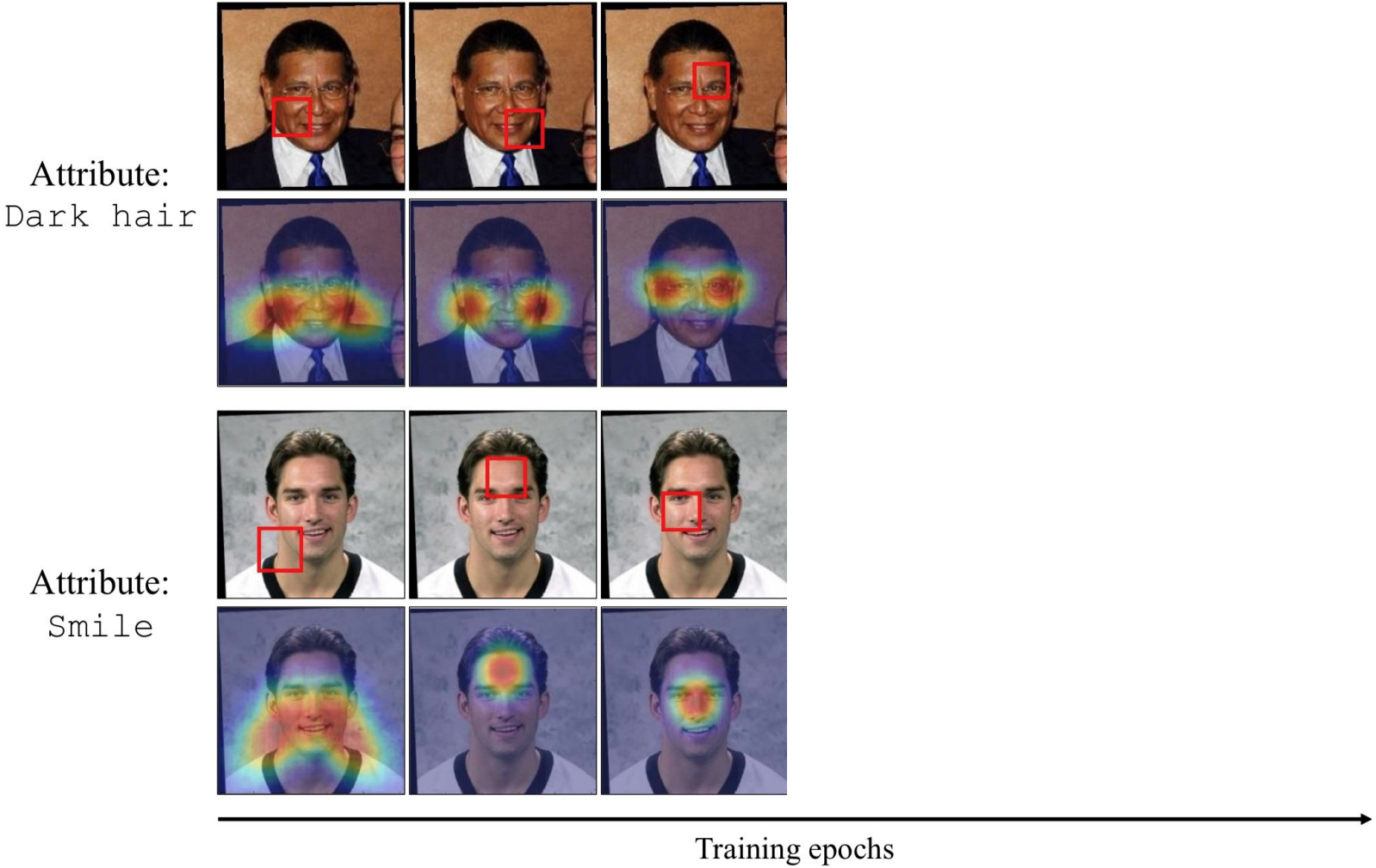
Attribute:
Dark hair

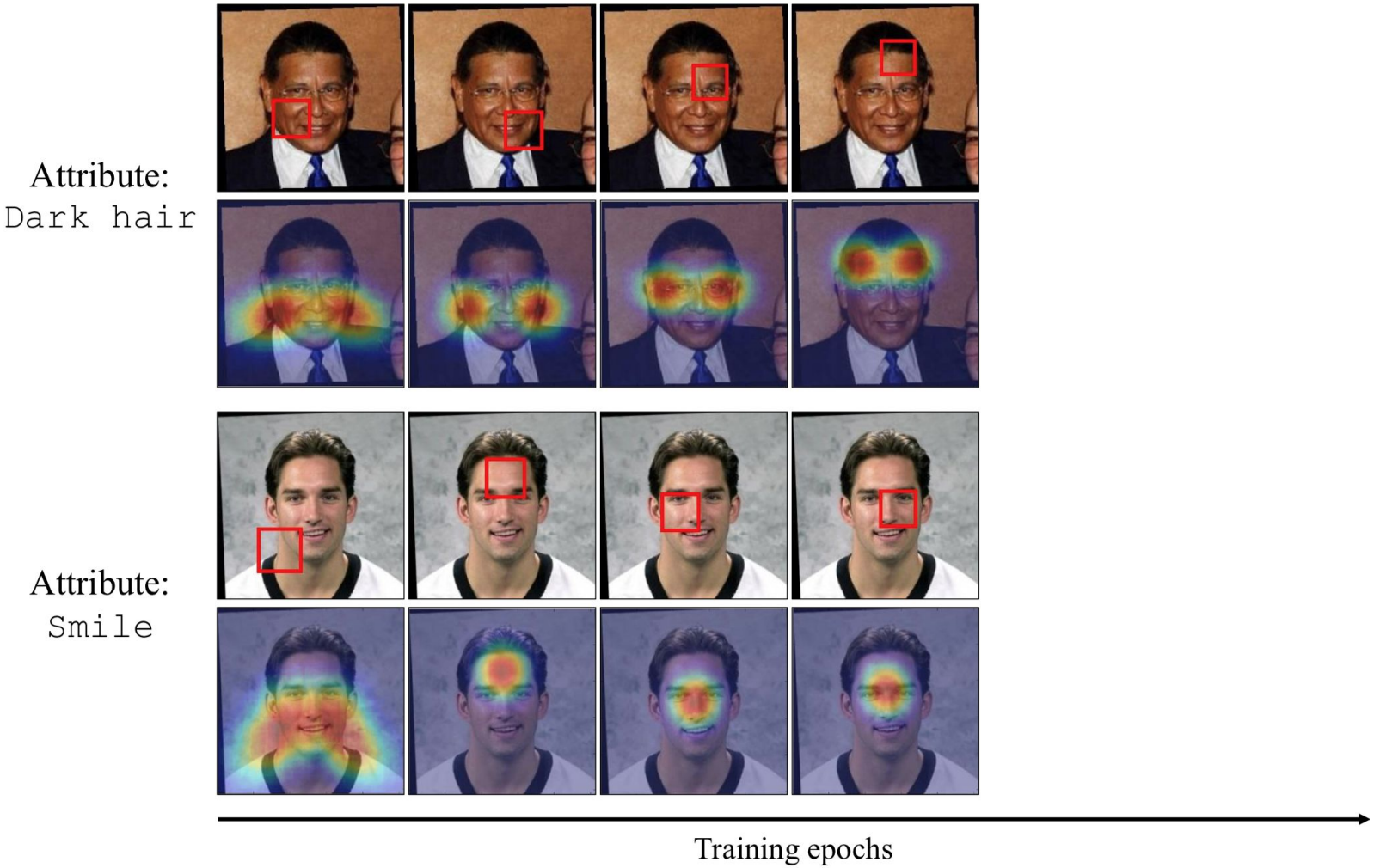


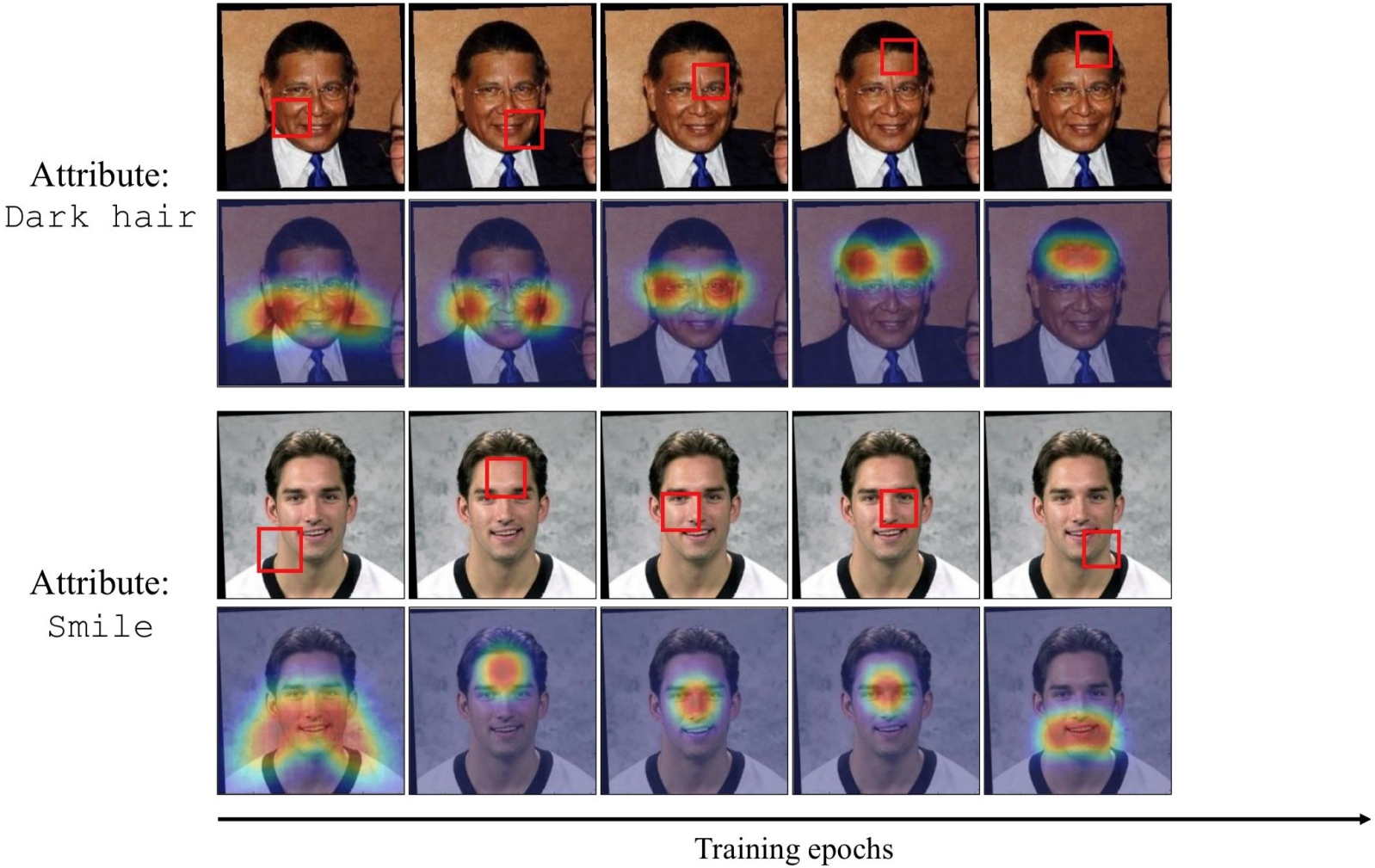
Attribute:
Smile

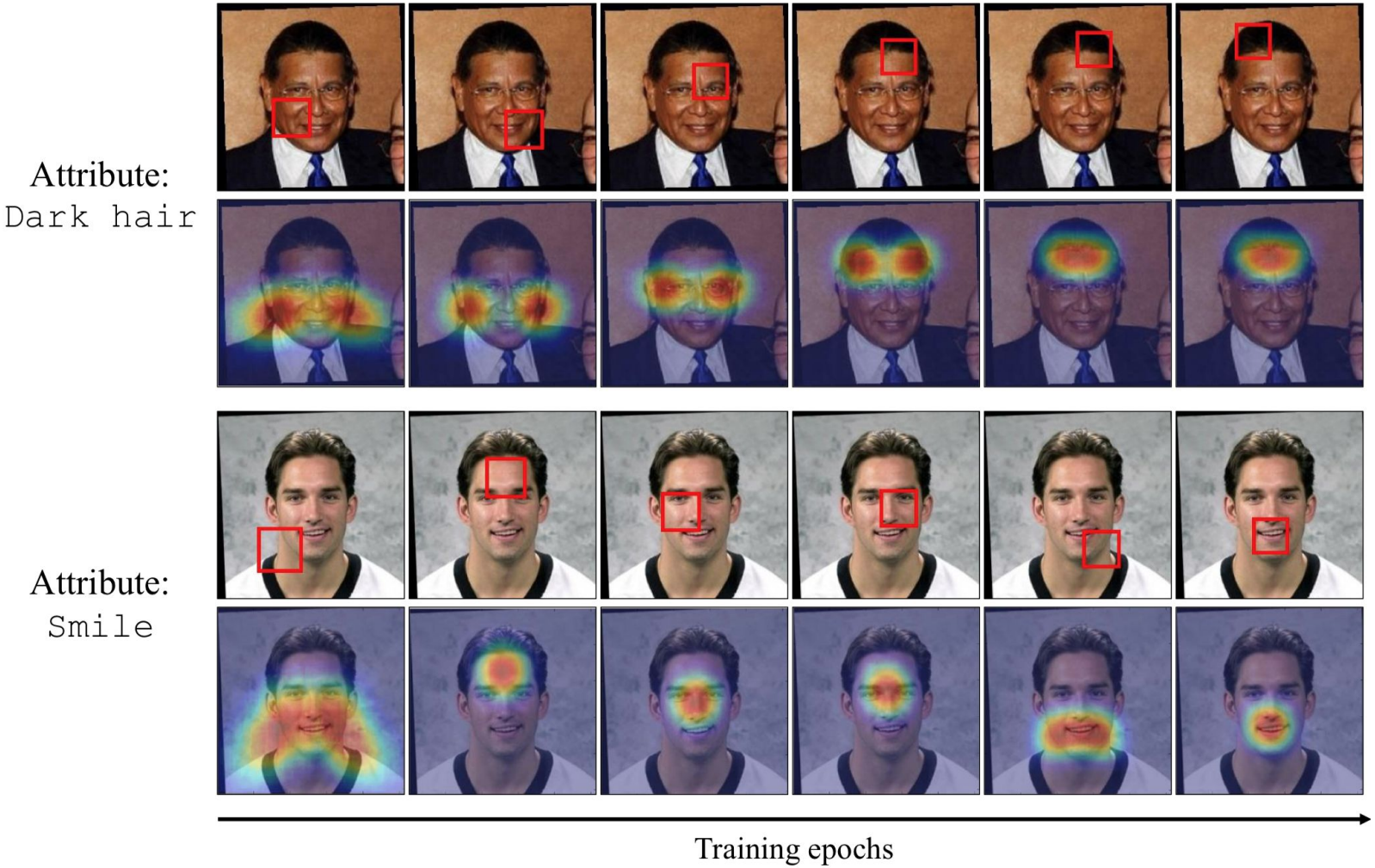


Training epochs


Attribute:
Dark hair
Attribute:
Smile
Training epochs


Attribute:
Dark hair
Attribute:
Smile
Training epochs


Attribute:
Dark hair
Attribute:
Smile
Training epochs


Attribute:
Dark hair
Attribute:
Smile
Training epochs

Unique Datasets

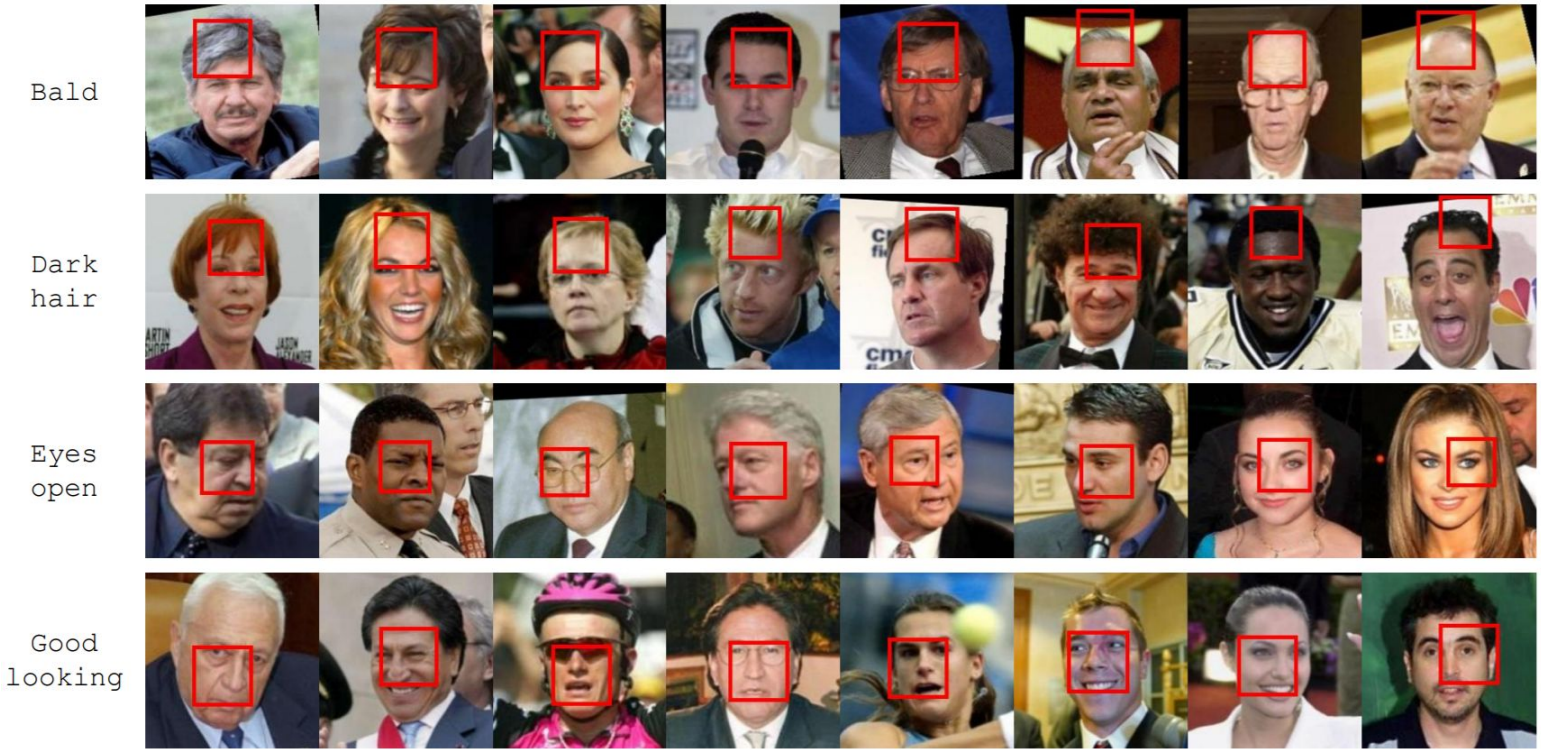
Contributions

End-to-end network that simultaneously performs attribute ranking and localization.

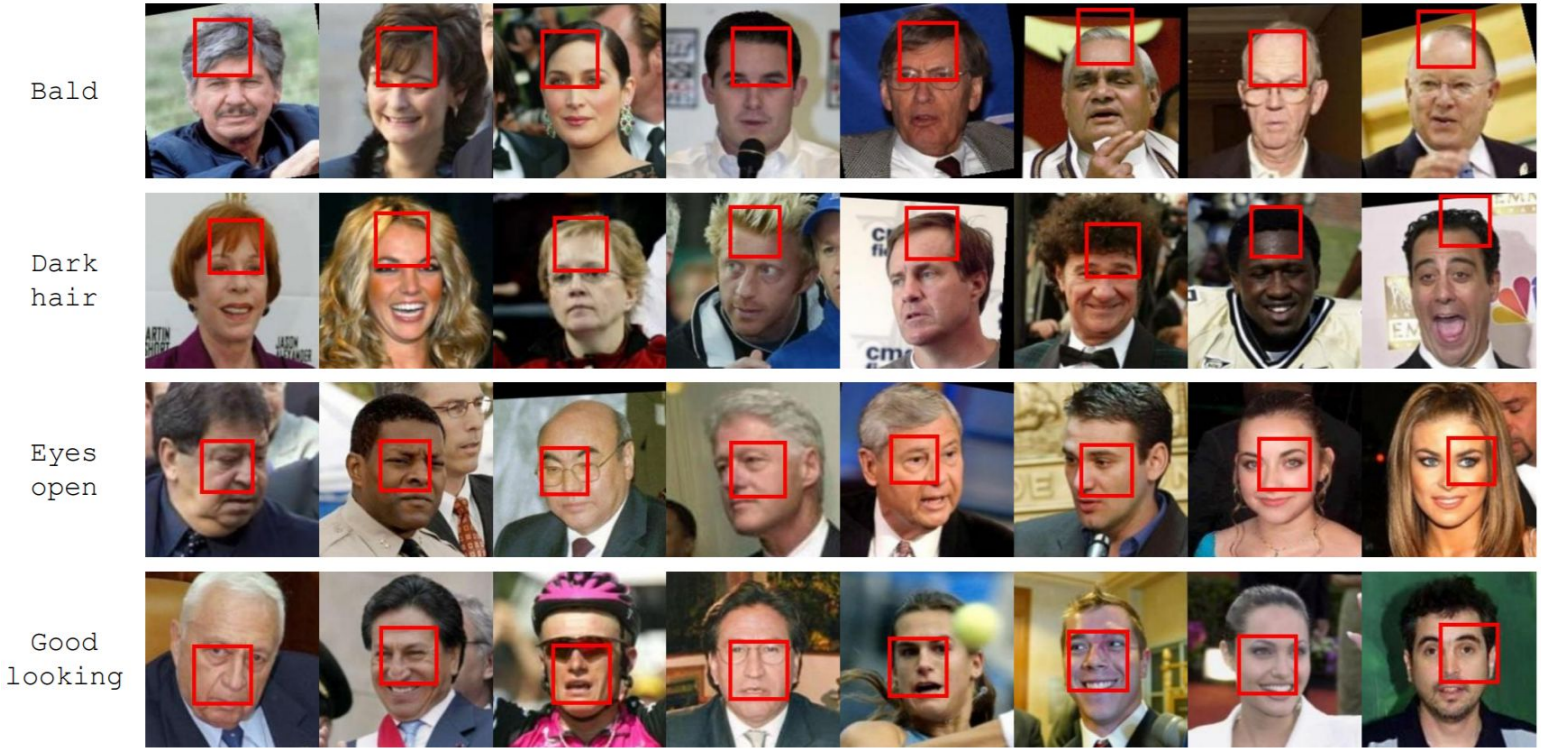
- Leverages Siamese network for ranking
- Spatial transformer localizes relevant image regions
- Generalizable; tested on face, shoe, and outdoor datasets

Torch Walkthrough

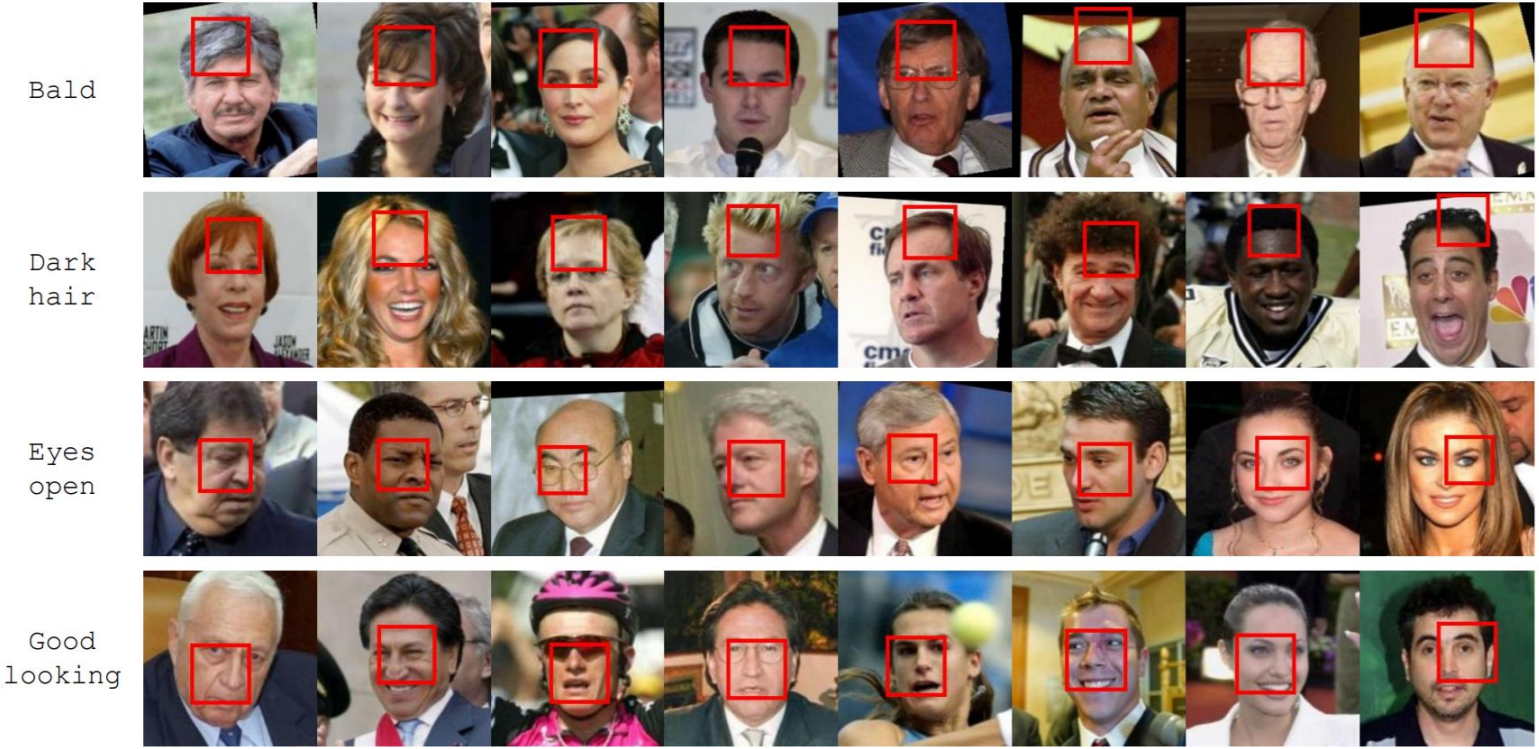
Experiments



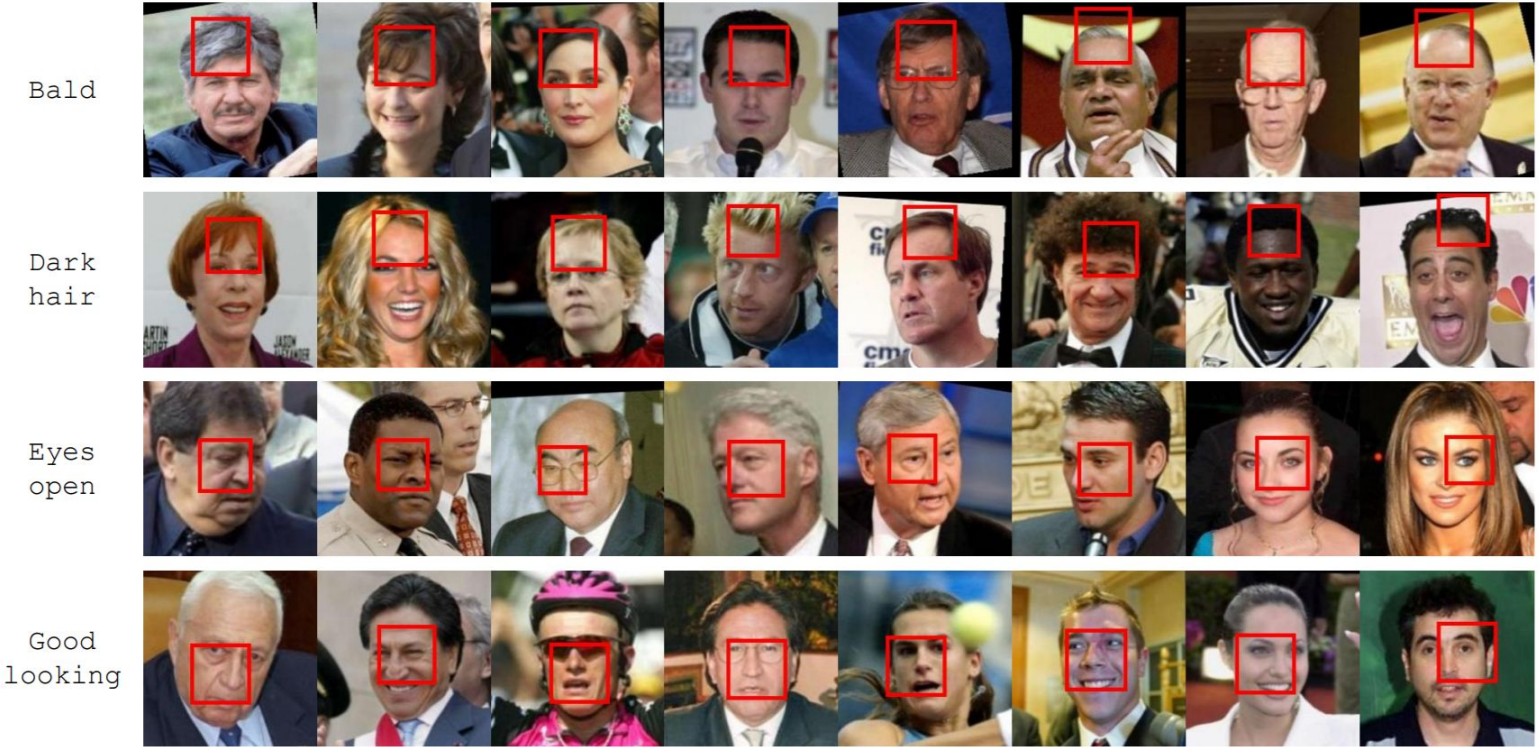
Even though we learn the scale of the STN, we can see that the size of the boxes is almost the same.



The size is of the box is close to ⅓, the initialized scale.



Since the scale doesn’t change much from the initialization scale, we try different initializations



The bounding box for bald head seems not to cover the whole related region

initialization scale
=0.66

Bald Head

initialization scale	STN output accuracy	Combined model accuracy
0.33	0.759	0.788
0.66	0.832	0.828

Bald Head

initialization scale	STN output accuracy	Combined model accuracy
0.33	0.759	0.788
0.66	0.832	0.828

Significant improvement on performance by changing initialization scale

Eyes open

scale=0.33
0.923

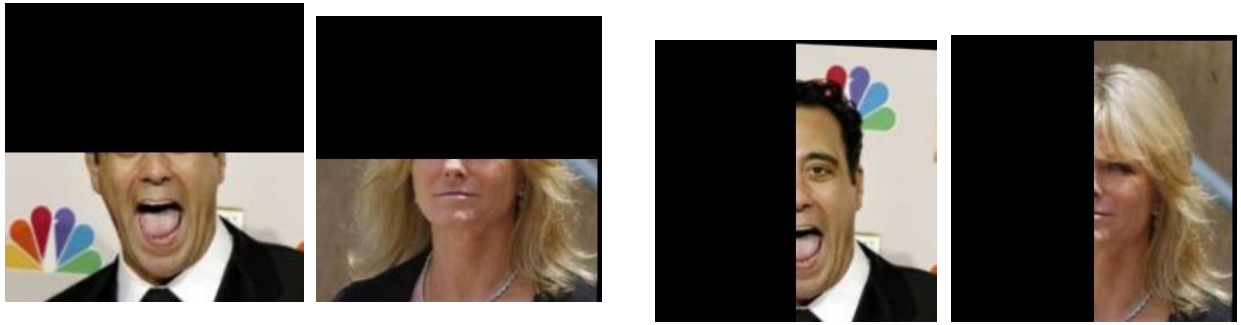
scale=0.17
0.919

Since is not learning the network is not learning to change the initialized scale, we increase the learning rate of scale

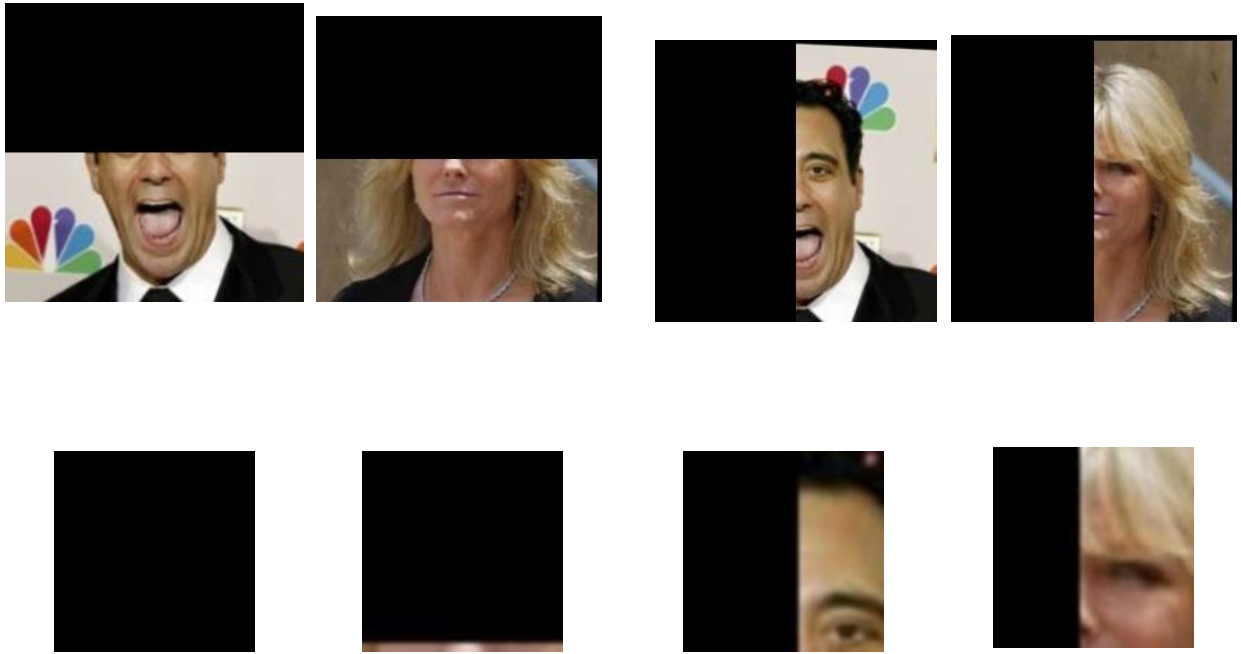
Modification	STN output accuracy
initialization scale 0.33	0.759
initialization scale 0.66	0.832
initialization scale 0.66, Scale learning rate x10	0.810
initialization scale 0.33, Scale learning rate x100	0.777

What does the localization network actually learn?

Bald head



Bald head



Eyes open

Mouth open

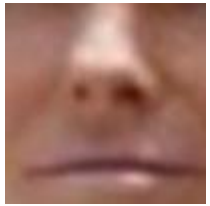
Eyes open

Mouth open

Mouth open

Bald head

Mouth open

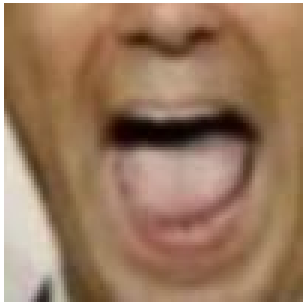


Bald head

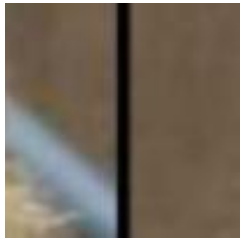
Bald head

Eyes open

Bald head



Eyes open



Thank you!