DeepGender: Occlusion and Low Resolution Robust Facial Gender Classification via Progressively Trained Convolutional Neural Networks with Attention

Felix Juefei Xu, Eshan Verma*, Parag Goel, Anisha Cherodian, and Marios Savvides*

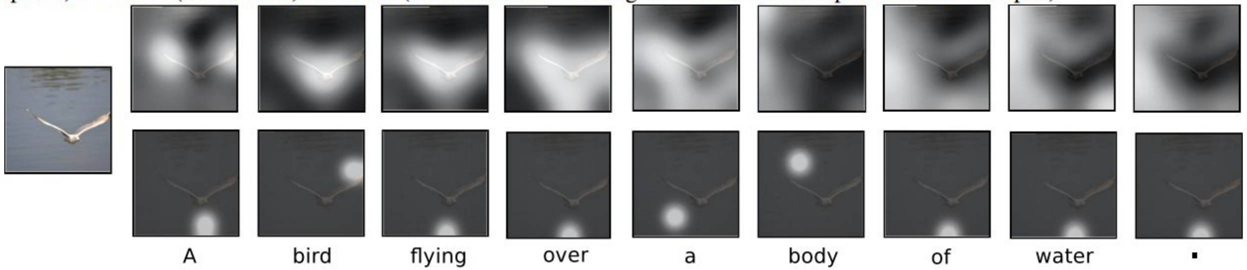
*Authors contribute equally.

CVPR 2016

Carnegie Mellon University
CyLab Biometrics Center

Motivation

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)



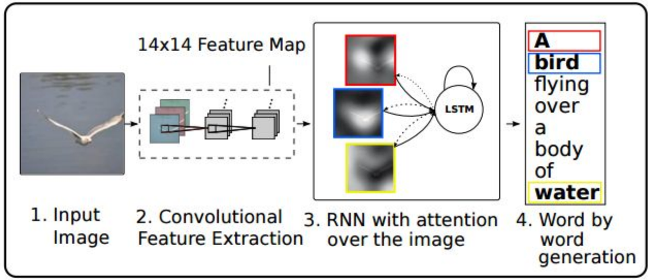
Attention based model [1]:

Encoder: CNN as feature extractor.

Decoder: RNN w/ LSTM, learns attention mechanism.

Visualization: the network can automatically fix its gaze on the salient objects (regions) in the image while generating the image caption word by word.

Figure 1. Our model learns a words/image alignment. The visualized attentional maps (3) are explained in section 3.1 & 5.4



Q: Can we control/enforce the attention shift in CNN?

[1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell - neural image caption generation with visual attention", ICML 2015.

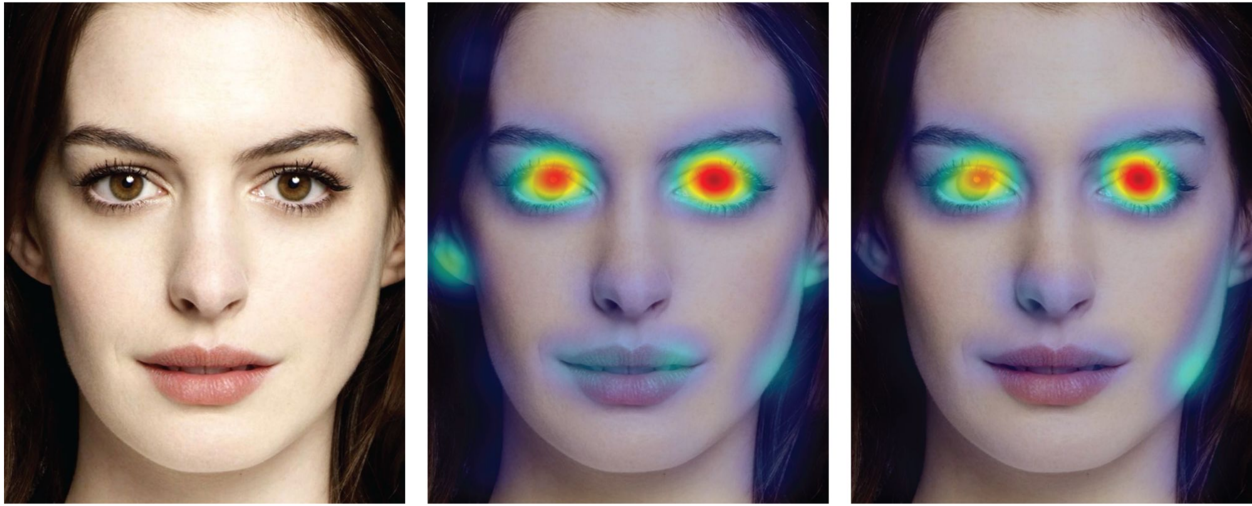
Carnegie Mellon University
CyLab Biometrics Center

Motivation

From previous work, we know that the periocular region provides the most important cues for determining gender information.

The periocular region is also the most salient region on human faces.

Input Image | *Image Signature - LAB* | *Image Signature - RGB*

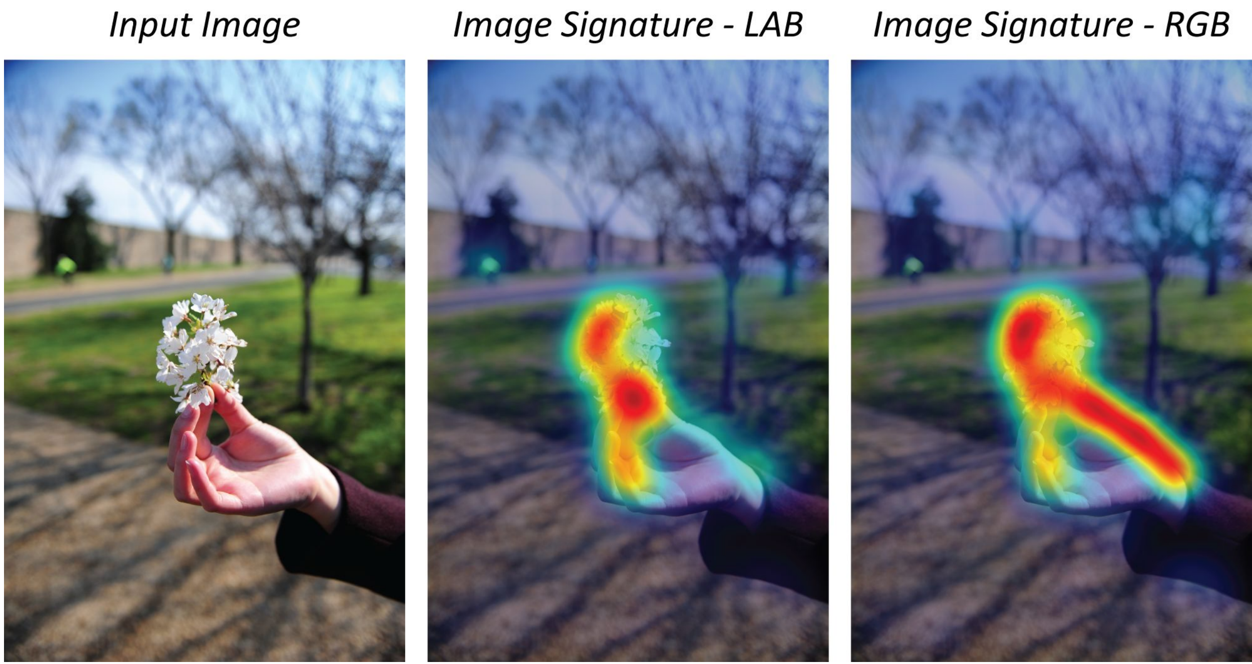


Motivation

Q: How can we let the CNN shift its attention towards the periocular region, where gender classification has been proven to be the most effective?

The answer comes from our day-to-day experience with photography.

The sharp foreground object attracts the most attention in the saliency heat map.



As opposed to the blurred / out-of-focus background content.

Motivation

We should be able to answer these following questions first before designing the progressive training paradigm.

Q: How can we let the CNN shift its attention towards the periocular region, where gender classification has been proven to be the most effective? (previous slide)

Q: Why not just use the periocular region crop?

Q: Why blurring instead of blackening out?

Q: Why not let CNN directly learn the blurring step?

Motivation

Q: Why not just use the periocular region crop?

Although periocular region is the best for gender classification, we still want to resort to other facial parts (beard/moustache) for providing valuable gender cues. Especially true when periocular region is less ideal (sunglasses).

To strike a good balance between full face-only and periocular-only models, we carry out a progressive training paradigm for CNN that starts with the full face, and progressively zoom into the periocular region by leaving other facial regions blurred.

Hope the network is sufficiently generalized.

Motivation

Q: Why blurring instead of blackening out?

We just want to steer the focus, rather than completely eliminate the background. Blackening would create abrupt edges.

When blurred, low frequency information is still well preserved. One can still recognize the content of the image, e.g., dog, human face, objects, etc. from a blurred image.

Blurring outside the periocular region, and leaving the high frequency details at the periocular region will both help providing global and structural context of the image, as well as keeping the minute details intact at the region of interest.

Motivation

Q: Why not let CNN directly learn the blurring step?

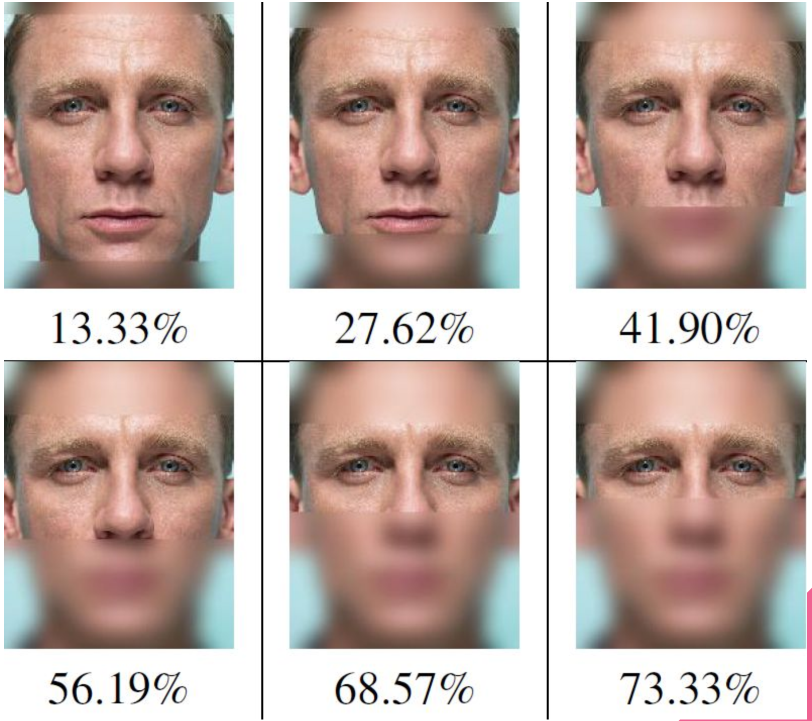
CNN filters operate on the entire image, and blurring only part of the image is a pixel location dependent operation and thus is difficult to emulate in the CNN framework.

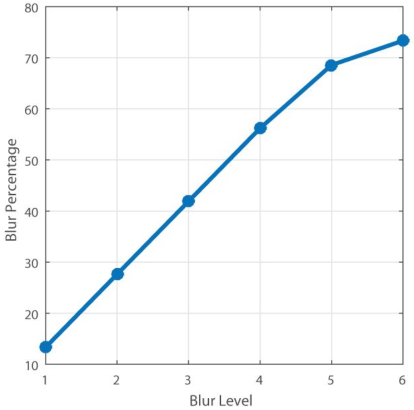
Enforcing Attention in the Training Images

We heuristically choose 7 blur levels, including the one with no blur at all.

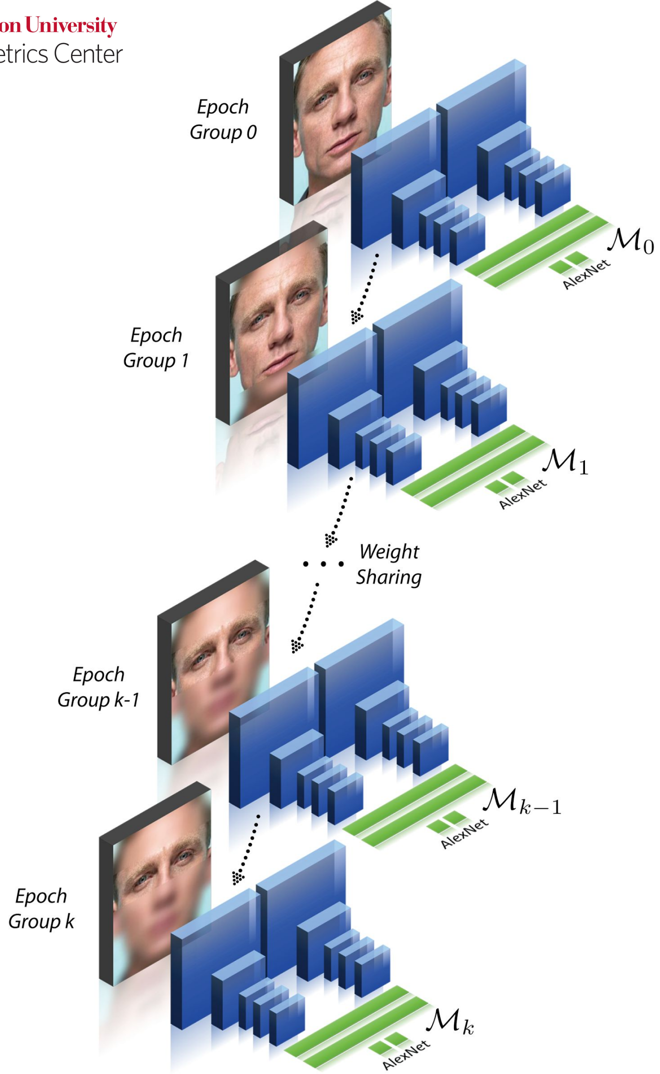
Gaussian blur kernel with σ = 7.

Doing this is conceptually enforcing the network attention in the training images without the need of changing the network architecture.





Carnegie Mellon University
CyLab Biometrics Center



Progressively Trained CNN with Attention

Training starts with the first epoch group (Epoch Group 0, images with no blur), and the first CNN model M_0 is obtained and frozen after convergence.

Then, we input the next epoch group for tuning the M_0 and in the end produce the second model M_1. Sequentially obtain models: M_1 to M_k.

AlexNet, 2-way softmax.

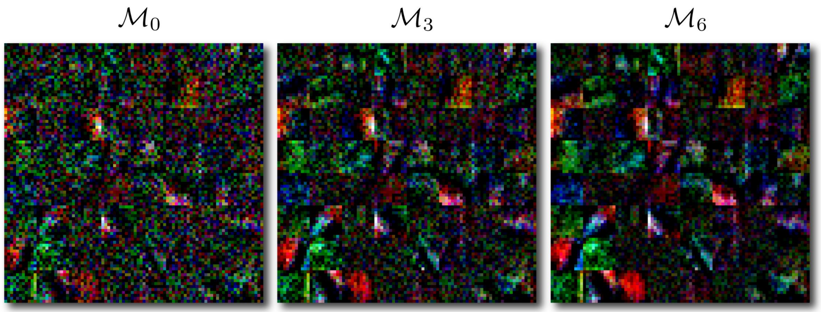
Each M_j (j=0, …, k) is trained with 1000 epochs, with a batchsize of 128.

Implicit Low-Rank Regularization in CNN

We have shown that the low-pass filtering in Fourier analysis is closely related to the low-rank approximation in SVD.

In the context of this work, progressively training the CNN using blurred images serves as an implicit low-rank regularizer.

This phenomenon is loosely observed through the visualization of the trained filters, which will be further analyzed and studied in future work.



Database

Training set: sourced from 5 different datasets. (Table 2)

Dimension: 168x210

Testing set: Pinellas County Sheriff's Office (PCSO) database, we use 400K out of 1.4M. To be added occlusion and low-res degradations.

Dimension: 168x210

Table 2: Datasets used for progressive CNN training.

DB Name	Males	Females
JNET	1900	1371
mugshotDB	1772	805
Pinellas Subset	13215	3394
pdx2	46346	12402
olympic2012	4164	3634
Total	67397	21606
	89003	

Pre-processing on 400K PCSO Testing Images

Row 1: random missing pixel occlusions

Row 2: random additive Gaussian noise occlusions

Row 3: random contiguous occlusions

Percentage of degradation for Row 1-3: 10%, 25%, 35%, 50%, 65%, 75%.

Row 4: various zooming factors (2x, 4x, 8x, 16x) for low-resolution degradations

Experiment 1: Occlusion Robustness

Experiments on the 400K PCSO mugshot database (artificial occlusions)

(1) Random missing pixels occlusions
(2) Random additive Gaussian noise occlusions
(3) Random contiguous occlusions

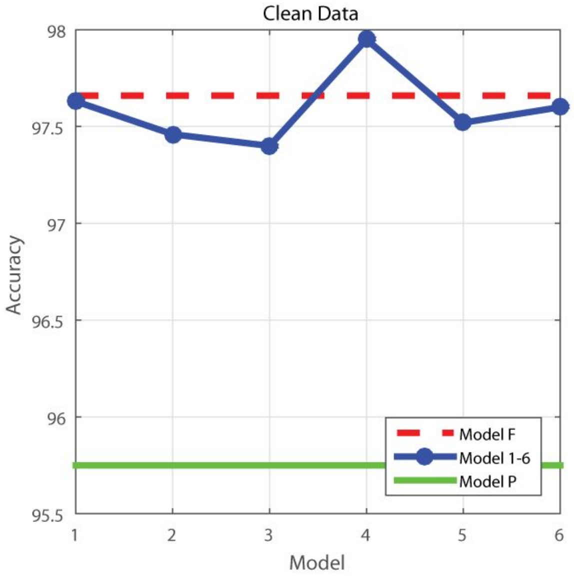
Baseline on Clean Images

This is the gender classification accuracies on the 400K PCSO database.

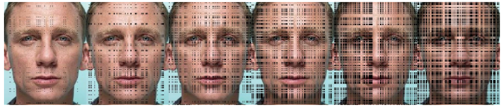
Images are clean, without artificially added degradations.

As expected, if the testing images are clean, it is preferable to use M_F rather than M_P.

M_F corresponds to the model trained on full face (equivalent to M_0), and M_P is one trained using only periocular region (last Epoch Group only). M_1-M_6 are the incremental models trained.



Carnegie Mellon University
CyLab Biometrics Center



(1) Random Missing Pixels Occlusions

M_5 performs the best with M_6 showing a dip, suggesting a tighter periocular region is not well-suited for such application.

Notice a flip in performance of M_F and M_P going from the 10% to 25% with the periocular model generalizing better for higher corruptions. The trend of improving performance between progressively trained models is maintained.

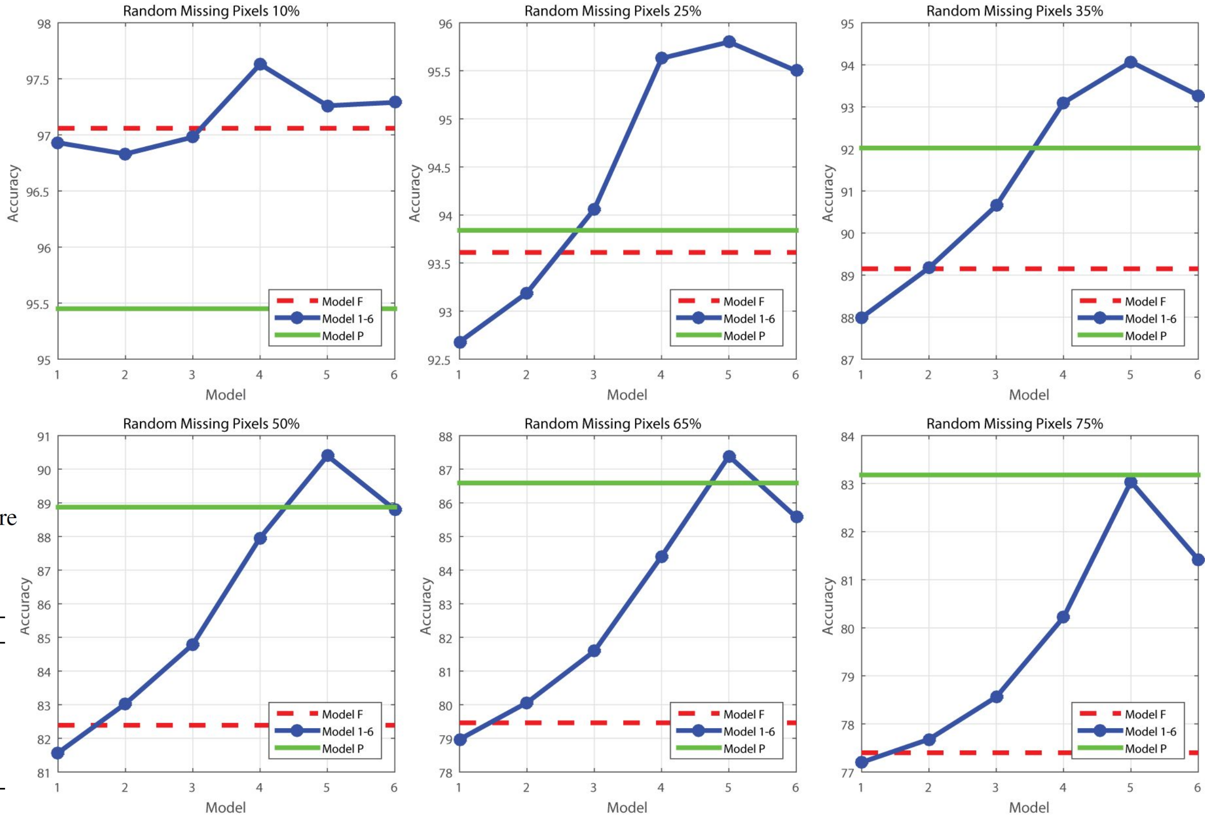


Table 3: Overall classification accuracy on the PCSO (400K). Images are corrupted with **random missing pixels** of various percentages.

Corrup.	0%	10%	25%	35%	50%	65%	75%
$\mathcal{M}_F$	97.66	97.06	93.61	89.15	82.39	79.46	77.4
$\mathcal{M}_1$	97.63	96.93	92.68	87.99	81.57	78.97	77.2
$\mathcal{M}_2$	97.46	96.83	93.19	89.17	83.03	80.06	77.68
$\mathcal{M}_3$	97.4	96.98	94.06	90.65	84.79	81.59	78.56
$\mathcal{M}_4$	97.95	97.63	95.63	93.1	87.96	84.41	80.22
$\mathcal{M}_5$	97.52	97.26	95.8	94.07	90.4	87.39	83.04
$\mathcal{M}_6$	97.6	97.29	95.5	93.27	88.8	85.57	81.42
$\mathcal{M}_P$	95.75	95.45	93.84	92.02	88.87	86.59	83.18

Carnegie Mellon University
CyLab Biometrics Center

(2) Random Additive Gaussian Noise Occlusions

M_4 - M_6 perform best for medium noise. For high noise, M_5 is the most robust.

Just as before, as the noise increases, the trend undertaken by the performance of M_F & M_P and M_5 & M_6 is maintained and so is the performance trend of the progressively trained models.

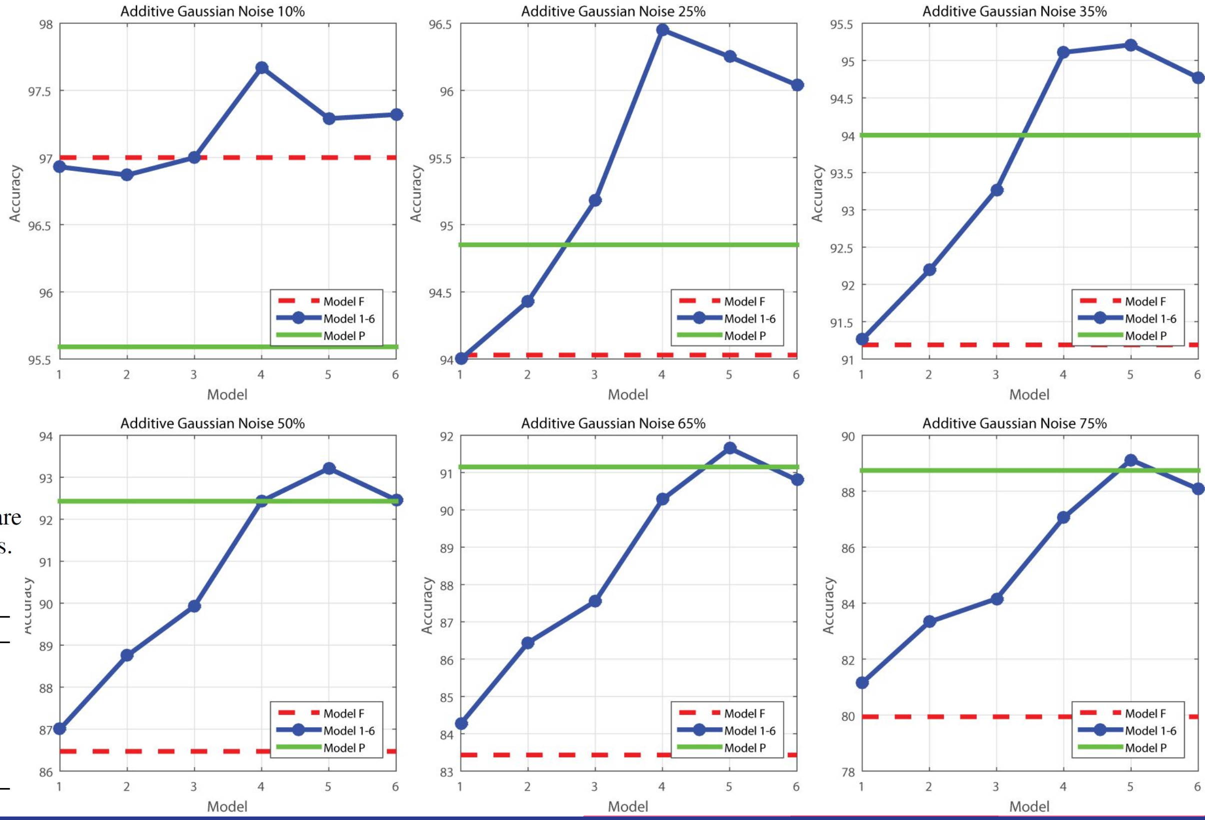


Table 4: Overall classification accuracy on the PCSO (400K). Images are corrupted with **additive Gaussian random noise** of various percentages.

Corrup.	0%	10%	25%	35%	50%	65%	75%
$\mathcal{M}_F$	97.66	97	94.03	91.19	86.47	83.43	79.94
$\mathcal{M}_1$	97.63	96.93	94	91.26	87	84.27	81.15
$\mathcal{M}_2$	97.46	96.87	94.43	92.19	88.75	86.44	83.33
$\mathcal{M}_3$	97.4	97	95.18	93.27	89.93	87.55	84.16
$\mathcal{M}_4$	97.95	97.67	96.45	95.11	92.43	90.28	87.06
$\mathcal{M}_5$	97.52	97.29	96.25	95.21	93.21	91.65	89.12
$\mathcal{M}_6$	97.6	97.32	96.04	94.77	92.46	90.8	88.08
$\mathcal{M}_P$	95.75	95.59	94.85	94	92.43	91.15	88.74

(3) Random Contiguous Occlusions

The most realistic occlusions are the first few cases, others are extreme cases.

For the former cases, M_1 - M_3 are able to predict the classes with the highest accuracy.

Our scheme of focused saliency helps generalizing over occlusions.

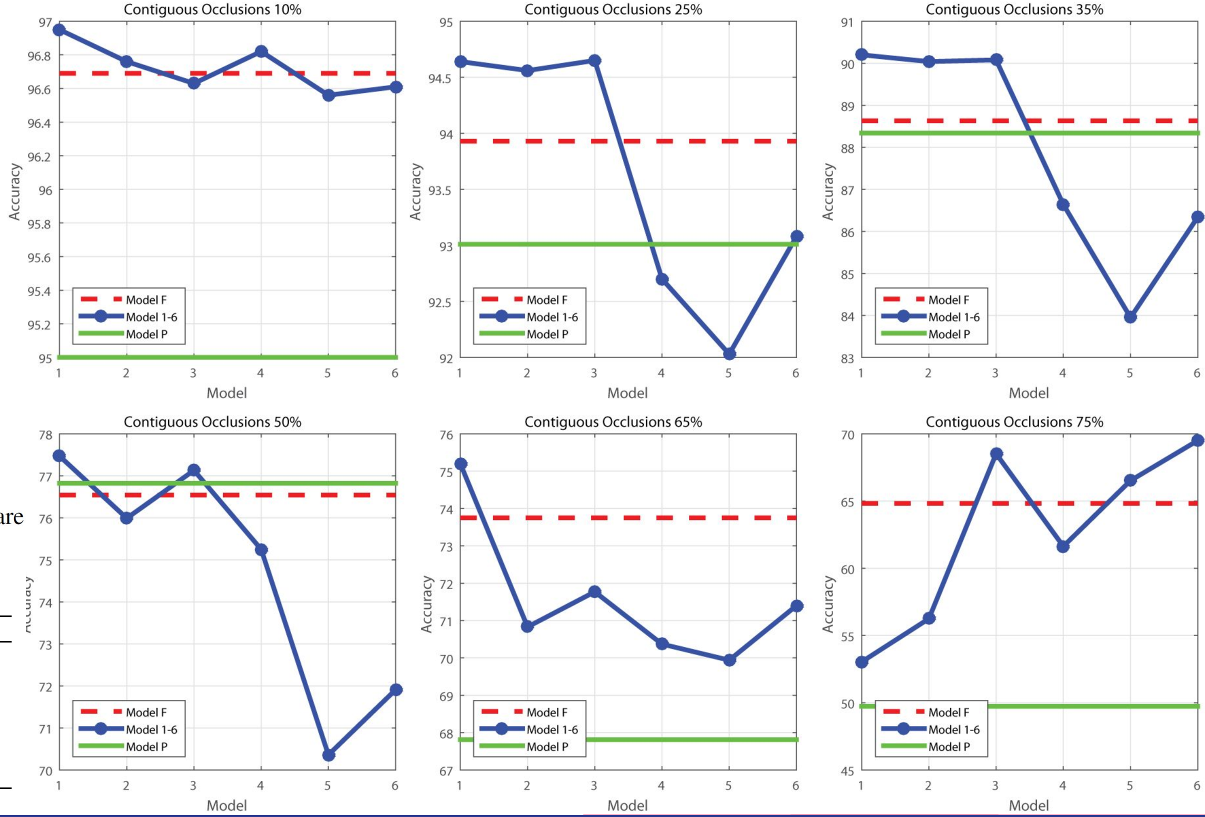


Table 5: Overall classification accuracy on the PCSO (400K). Images are corrupted with **random contiguous occlusions** of various percentages.

Corrup.	0%	10%	25%	35%	50%	65%	75%
$\mathcal{M}_F$	97.66	96.69	93.93	88.63	76.54	73.75	64.82
$\mathcal{M}_1$	97.63	96.95	94.64	90.2	77.47	75.2	53.04
$\mathcal{M}_2$	97.46	96.76	94.56	90.04	75.99	70.83	56.25
$\mathcal{M}_3$	97.4	96.63	94.65	90.08	77.13	71.77	68.52
$\mathcal{M}_4$	97.95	96.82	92.7	86.64	75.25	70.37	61.63
$\mathcal{M}_5$	97.52	96.56	92.03	83.95	70.36	69.94	66.52
$\mathcal{M}_6$	97.6	96.61	93.08	86.34	71.91	71.4	69.5
$\mathcal{M}_P$	95.75	95	93.01	88.34	76.82	67.81	49.73

Experiment 2: Low Resolution Robustness

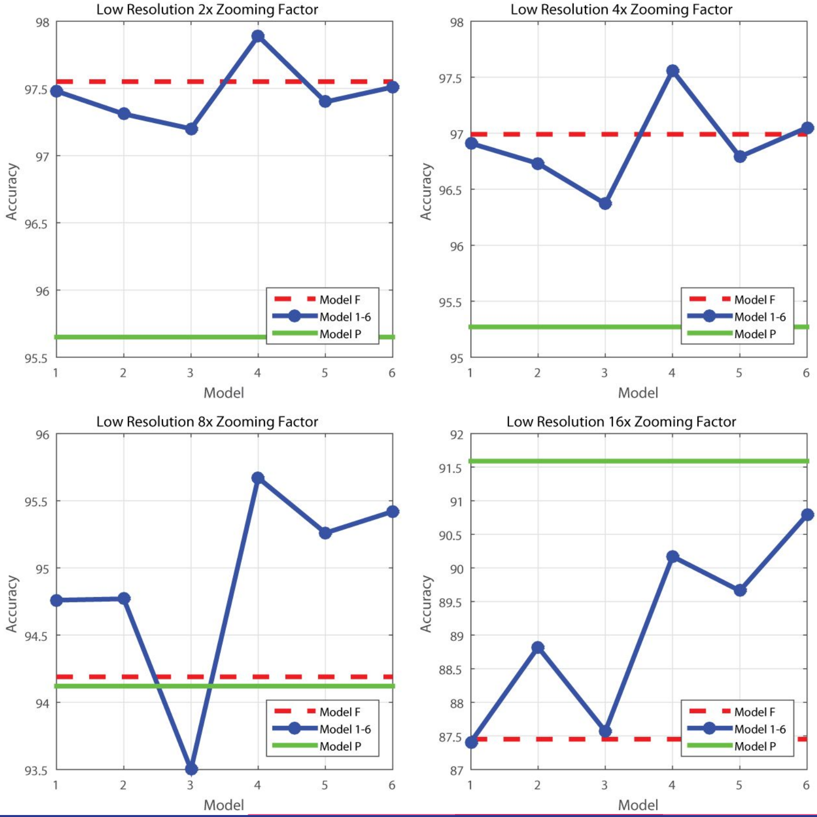
Experiments are on the 400K PCSO mugshot database.

For cases 2x, 4x, and 8x, the trend between M_1-M_6 and their performance with respect to M_F is maintained.

For 16x case, progressive models M_1-M_6 are still better than full face model M_F.

Table 7: Overall classification accuracy on the PCSO (400K). Images are down-sampled to a **lower resolution** with various zooming factors.

Zooming Factor	1x	2x	4x	8x	16x
$\mathcal{M}_F$	97.66	97.55	96.99	94.19	87.45
$\mathcal{M}_1$	97.63	97.48	96.91	94.76	87.41
$\mathcal{M}_2$	97.46	97.31	96.73	94.77	88.82
$\mathcal{M}_3$	97.4	97.2	96.37	93.5	87.57
$\mathcal{M}_4$	97.95	97.89	97.56	95.67	90.17
$\mathcal{M}_5$	97.52	97.4	96.79	95.26	89.66
$\mathcal{M}_6$	97.6	97.51	97.05	95.42	90.79
$\mathcal{M}_P$	95.75	95.65	95.27	94.12	91.59



Conclusion and Discussion

The intuition:

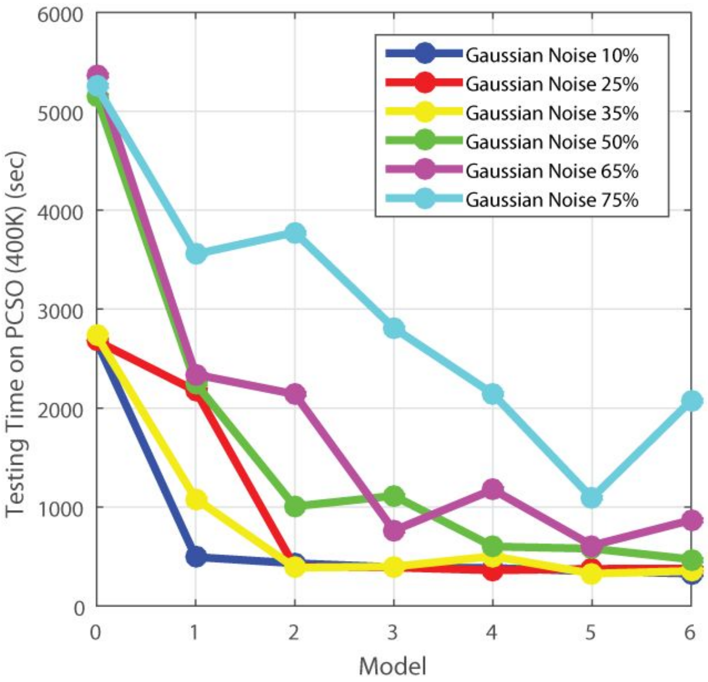
(1) To have the network focus on the periocular region of the face for gender classification.

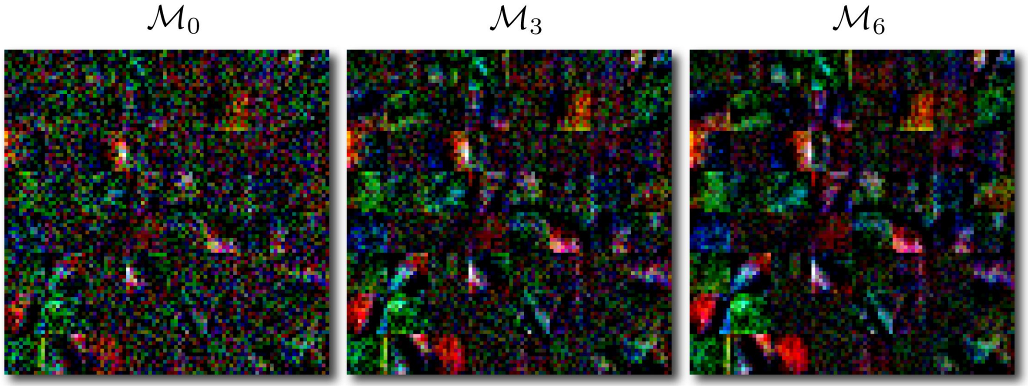
(2) To preserve contextual information of facial contours to generalize better over occlusions.

Our hypothesis is indeed true and that for a given occlusion set, it is possible to have high accuracy from a model that encompasses both of above stated properties.

We did not train on any occluded data, or optimize for a particular type of occlusions, our models can generalize well.

Future Work





We have observed significant testing time savings from M0 to M6. We have thus visualize the learned filters. It seems that after progressive training, the filters are smoother, and we will study the connection between the two in the future.

Thank you!
Questions?

Check out the poster.