Yingyan Shi

yyshi17@fudan.edu.cn

Institute of Brain-Inspired Circuits and Systems

Fudan University

Facial Attribute Analysis

---

1. A Survey to Deep Facial Attribute Analysis

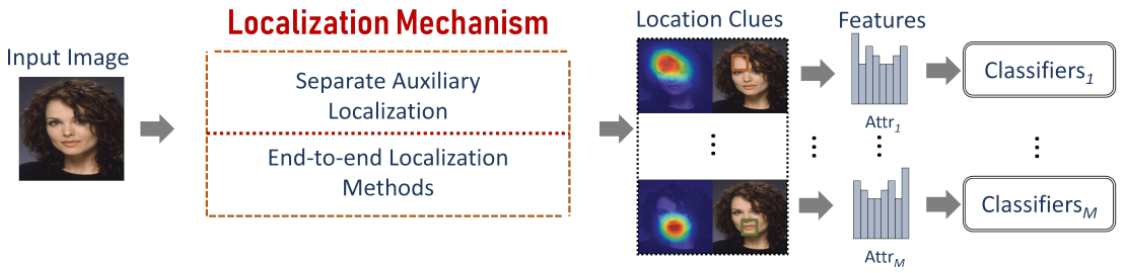
arXiv:1812

赫然

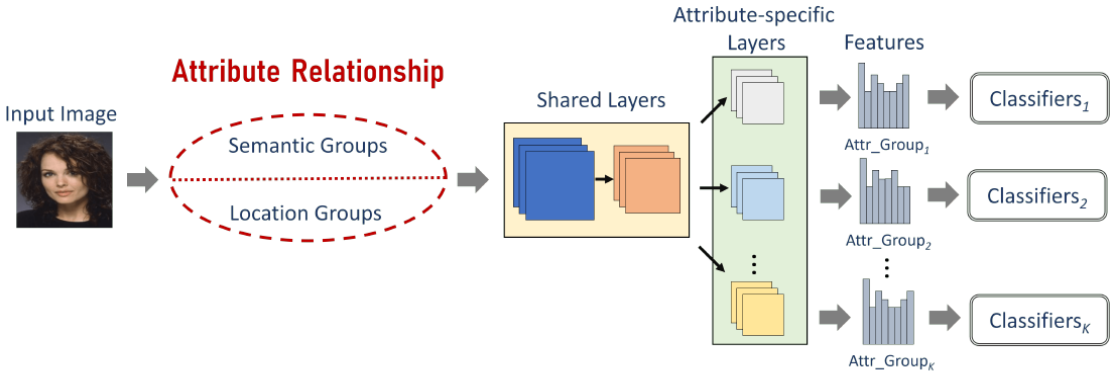
Facial attribute analysis:

- facial attribute estimation
  - Part-based methods (pays more attention to locate attributes)
    - Separate auxiliary localization
      - the localization and estimation are operated in a separate and independent manner (detection then classification)
    - end-to-end localization
      - exploit the locations of facial attributes and predict their presences simultaneously in end-to-end frameworks (general object detection)
  - Holistic methods (focus on more on modeling attribute relationships)
    - without any extra localization modules

- modeling the association and distinction among different attributes to explore the complementary information, by designing various networks with sharing features from different layers.



Part-based Deep FAE Method



Holistic Deep FAE Method

The key of modeling attribute relationships is learning common features at low-level shared layers and exploring attribute-specific features at high level separated layers, where each separated layer corresponds to an attribute group.

- facial attribute manipulation

Data augmentation

2. AFFACT: Alignment-Free Facial Attribute Classification Technique

University of Colorado

2017 IEEE International Joint Conference on Biometrics (IJCB)

Explore:

- face alignment or
- Alignment-Free Facial Attribute Classification Technique ?
- 人脸对齐可以减少面部平面内的旋转，但是要求检测人脸关键点
- 测试时不用检测人脸关键点的话，就要在训练时使用大量的数据增强技术

ResNet, Alignment-Free Facial Attribute Classification Technique (AFFACT) with data augmentation.

only the detected bounding boxes rather than requiring alignment based on automatically detected facial landmarks

3.2. Data Augmentation for Training

While making use of the natural alignment of faces, we extend traditional data augmentation techniques to incorporate scaling, rotation, shifting, and blurring of training images. Given an image annotated with attribute labels and some facial landmarks [18], we align the face after applying random modifications to the scale, the angle, and the location of the bounding box. Specifically, given the labels of the two eyes $\vec{t}_{e_r}, \vec{t}_{e_l}$ and the mouth corners $\vec{t}_{m_r}, \vec{t}_{m_l}$ with $\vec{t} = (x, y)^{\mathrm{T}}$, we compute the center of both and their respective distance d:

$$\vec{t}_e = \frac{\vec{t}_{e_r} + \vec{t}_{e_l}}{2}, \quad \vec{t}_m = \frac{\vec{t}_{m_r} + \vec{t}_{m_l}}{2}, \quad d = \|\vec{t}_e - \vec{t}_m\|\,, \quad (1)$$

then we estimate the bounding box coordinates (specifying left, top, right and bottom coordinates x_l, y_t, x_r, y_b):

$$\begin{aligned} x_l &= x_e - 0.5 \cdot w \quad & x_r &= x_e + 0.5 \cdot w \\ y_t &= y_e - 0.45 \cdot h \quad & y_b &= y_e + 0.55 \cdot h \end{aligned} \quad (2)$$

as well as the rotation angle obtained from the eye locations:

$$\alpha = \arctan \frac{y_{e_r} - y_{e_l}}{x_{e_r} - x_{e_l}}\,. \quad (3)$$

The height h and width w of the bounding box are estimated based on the mouth-eye-distance: $h = w = 5.5 \cdot d$.

We have implemented a new data layer in Caffe [10] that randomly perturbs the bounding boxes of images in each training epoch *on the fly*, using Bob [1] to handle image transformations and OpenMP [4] for parallelization on the CPU. In this layer, an original image and its pre-computed bounding box and angle α are loaded, the scale $s = W/w = H/h$ is estimated, offsets for angle $r_\alpha \sim \mathcal{N}_{0,20}$, shift $r_y, r_x \sim \mathcal{N}_{0,0.05}$, and scale $r_s \sim \mathcal{N}_{1,0.1}$ are randomly drawn and added to the coordinates:

$$\begin{aligned} \tilde{x}_l &= x_l + \tilde{r}_x \cdot w, \quad & \tilde{x}_r &= x_r + \tilde{r}_x \cdot w, \quad & \tilde{\alpha} &= \alpha + r_\alpha, \\ \tilde{y}_t &= y_t + \tilde{r}_y \cdot w, \quad & \tilde{y}_b &= y_b + \tilde{r}_y \cdot w, \quad & \tilde{s} &= s \cdot \tilde{r}_s. \end{aligned} \quad (4)$$

Using these coordinates, the image is rotated, scaled, and cropped into an RGB image with resolution $W = H = 224$, and horizontally flipped with 50 % probability. To emulate smaller image resolutions, yielding blurred upscaled images, a Gaussian filter with a random standard deviation of $\sigma \sim \mathcal{N}_{0,3}$ is applied to smooth the image, which is finally fed as input to the network training. Fig. 2 shows some examples of training images cropped with and without random perturbations. These random perturbations are much larger than used in related work [25, 20, 23]. Note that all parameters have been selected based on the characteristics of the CelebA dataset and the facial attribute classification task. Particularly, the amount of random scale, angle, shift, and blur are selected to work with the rotations of the faces in the CelebA dataset, as well as with the face detector [24]. However, these parameters can easily be adapted for other image resolutions, contents, and face detectors.

3.3. Data Augmentation for Testing

Data augmentation can also effectively be used for testing in order to further enhance performances of DCNNs [7]. A common practice is to combine predictions of ten transformations into a final prediction – i.e., using their average – by rescaling the image and taking the center crop together with crops of the four corners of the original and the horizontally flipped images. In our experiments, we adopt this strategy and rescale the test images to a resolution of 256×256 pixels before taking those ten crops. We average the resulting DCNN output scores per attribute, and threshold the results at $\tau = 0$ to obtain the final attribute classification.

(a) Without Perturbations

(b) With Random Perturbations

Figure 2: RANDOMLY PERTURBED TRAINING IMAGES. The impact of our random perturbations to image alignment is shown. In (a) images are aligned using the corresponding bounding box and angle, which are computed based on hand-labeled facial landmarks, while (b) shows them perturbed with random scale, angle, shift, blur, and horizontal flip.

3. Real Time System for Facial Analysis

Finland

2017

https://github.com/mahehu/TUT-live-age-estimator

- a screen
- a webcam
- a computer (i7 2600 CPU only)

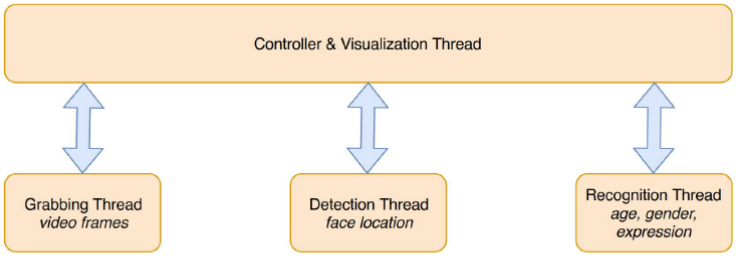


Figure 2. System architecture

Backbone: MobileNets: Efficient convolutional neural networks for mobile vision applications, 2017

Stage	*Network*	*Accuracy*
Detection	SSD-MobileNet, α=.75	67.2% (AP @0.5IoU)
Age	MobileNet, α=1.0	4.9 years (MAE)
Gender	MobileNet, α=1.0	88.3% (accuracy)
Expression	MobileNet, α=1.0	55.9% (accuracy)

4. ChaLearn Looking at People and Faces of the World: Face Analysis Workshop and Challenge 2016

CVPRW 2016

University of Nottingham

3 competitions:

- Age estimation
- Accessory classification
- Smile and gender classification

5. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks

IJCV 2016

ETH Zurich, Switzerland

face detector: Face detection without bells and whistles. ECCV 2014

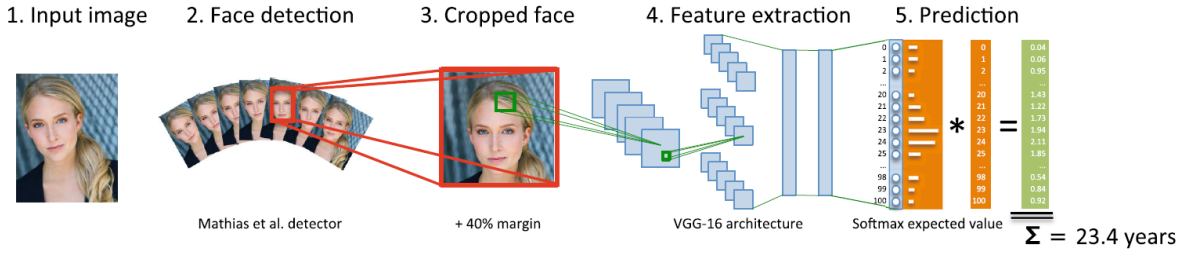
face alignment: the failure of the landmark detector is difficult to predict and harms the performance as it leads to wrong face alignments.

- We **explicitly handle rotation** by running the detector not only on the original image but on images rotated with steps of 5°
- the face with the highest detection score across all rotations is picked and then rotated to up frontal position

VGG-16 architecture

利用神经网络回归大范围数值时的输出策略，将输出的数值分成多个bin，尽量使各bin所包含的样本数量差不多，2 options:

- the predicted age is the age of the neuron with the highest probability
- the expected value over the SoftMax normalized output probabilities



6. Attributes for Improved Attributes: A Multi-Task Network Utilizing Implicit and Explicit Relationships for Facial Attribute Classification

AAAI 17

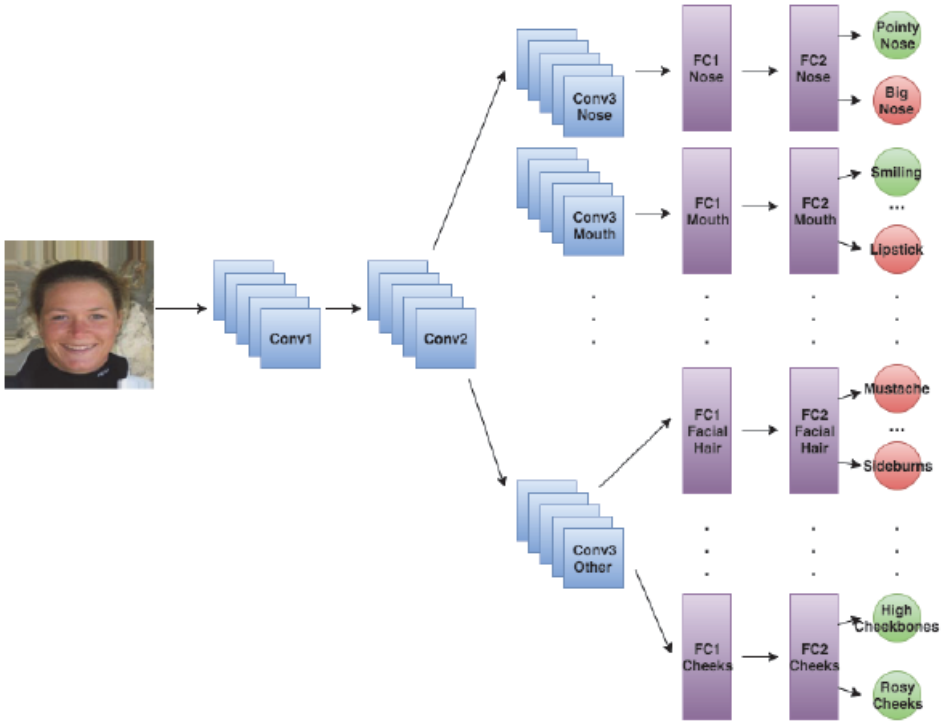
University of Maryland

a lot to be gained from shared information amongst attributes

framing the attribute prediction problem as a multi-task learning problem

Attribute relationships are learned implicitly at the lower levels, and explicitly in the higher grouped layers.

- by sharing lower layers of MCNN
- by grouping similar attributes in higher layers of MCNN
- by introducing an auxiliary network (AUX), which learns attribute relationships at the score level.



divides all the 40 attributes into 9 groups according to semantic

Group	Attributes
Gender	Male
Nose	Big Nose, Pointy Nose
Mouth	Big Lips, Lipstick, Mouth Slightly Open, Smiling
Eyes	Arched Eyebrows, Bags Under Eyes, Bushy Eyebrows, Eyeglasses, Narrow Eyes
Face	Attractive, Blurry, Heavy Makeup, Oval Face, Pale Skin, Young
AroundHead	Balding, Bangs, Black Hair, Blond Hair, Brown Hair, Earrings, Gray Hair, Hat, Necklace, Necktie, Receding Hairline, Straight Hair, Wavy Hair
FacialHair	5 o'clock Shadow, Goatee, Mustache, No Beard, Sideburns
Cheeks	High Cheekbones, Rosy Cheeks
Fat	Chubby, Double Chin

Table 1: Attributes and their corresponding groupings.

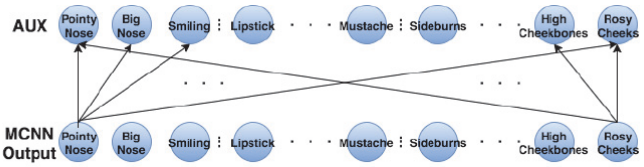


Figure 2: AUX network architecture. The output of the MCNN is fully connected to the final layer creating the 2-layer AUX network.

7. Multi-task Learning of Cascaded CNN for Facial Attribute Classification

arXiv: 1805

Xiamen University

exploit the inherent dependencies among these tasks:

1. face detection
2. facial landmark localization
3. facial attribute classification

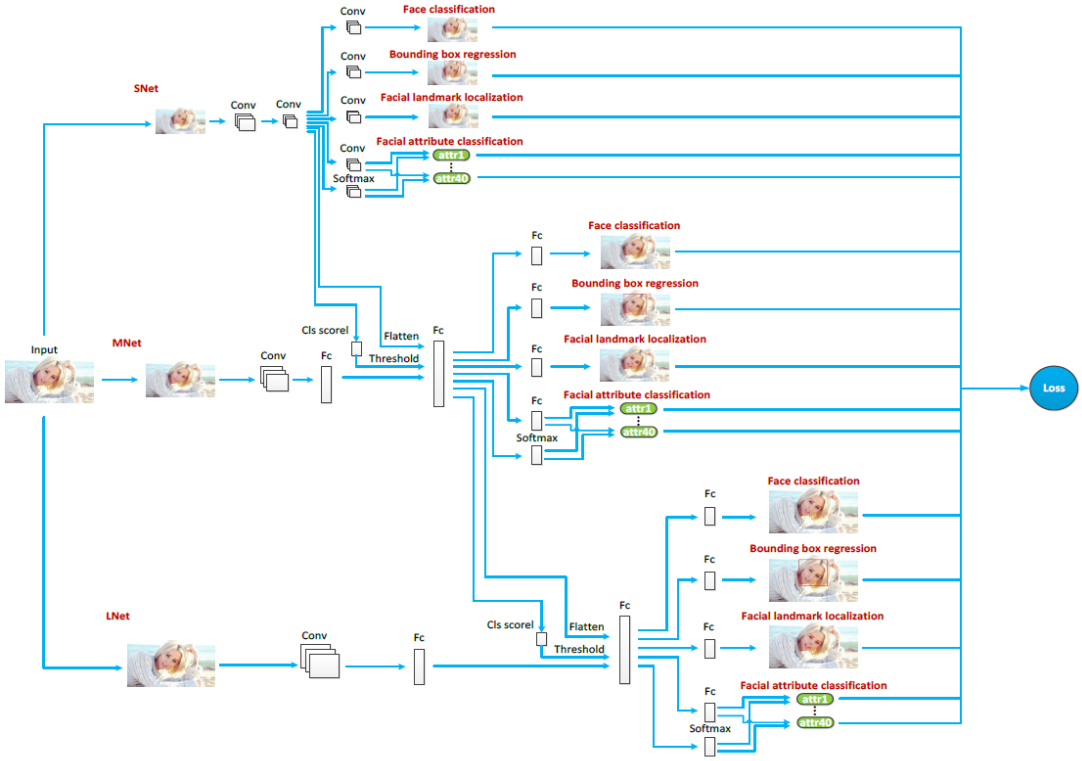


Fig. 2: The overall framework of the proposed MCFA method for FAC. The input image is resized to three different scales (i.e., 56×56, 112×112 and 224×224), which correspond to the inputs of the three cascaded sub-networks (i.e., S_Net, M_Net and L_Net).

Coarse-to-fine manner, end-to-end optimization:

- coarse feature: Small_Net
- fine feature: Middle_Net
- subtle feature: Large_Net

Backbone: VGG-16

8. A Deep Cascade Network for Unaligned Face Attribute Classification

arXiv: 1709

University of Maryland

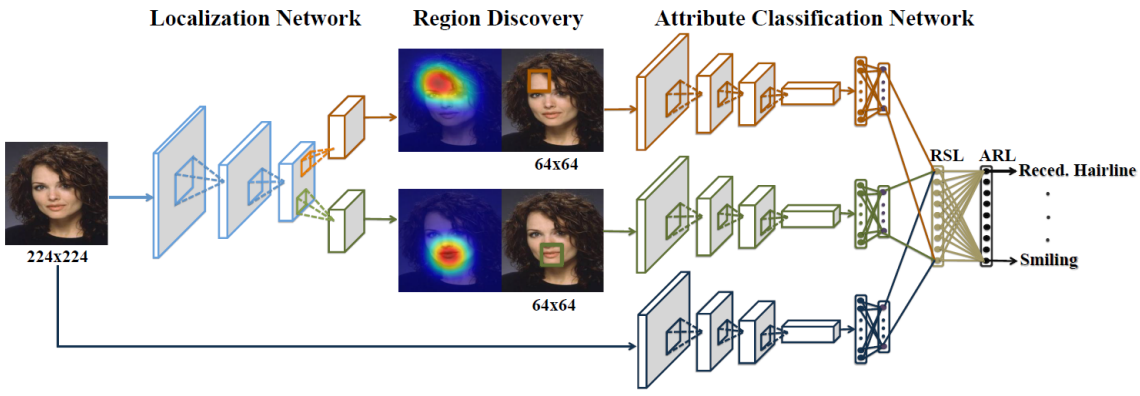


Figure 1: Overview of our face attribute recognition framework. It consists of a facial region localization (FRL) network and a Parts and Whole (PaW) classification network. The localization network detects a discriminative part for each attribute. Then the detected face regions and the whole face image are fed into the PaW classification network. The region switch layer (RSL) selects the relevant subnet for predicting the attribute, while the attribute relation layer (ARL) models the attribute relationships.

global average pooling for the localization task: GAP structure

multi-net learning: The idea is to simultaneously train the two different types of networks with the same attributes loss. This extra supervision from the classification branch regularizes the training process to search for a more discriminative solution.

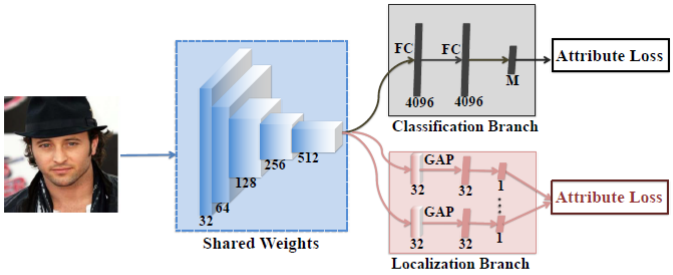


Figure 2: Multi-Net Learning.

将heatmap up-sample到原始大小，利用神经网络自动寻找属性相关的区域，即最大响应值的区域。

labeled regions unavailable, so weakly-supervised way where only face attribute labels are needed.

5 VGG-Net conv modules shared by all the attributes whose weights are initialized from the VGG-Face CNN.

- 40 part-based subnets
- 1 whole-image-based subnet

交替训练各组件，非端到端优化

9. Learning deep features for discriminative localization

CVPR 2016

周博磊，MIT CSAIL

https://github.com/metalbubble/CAM

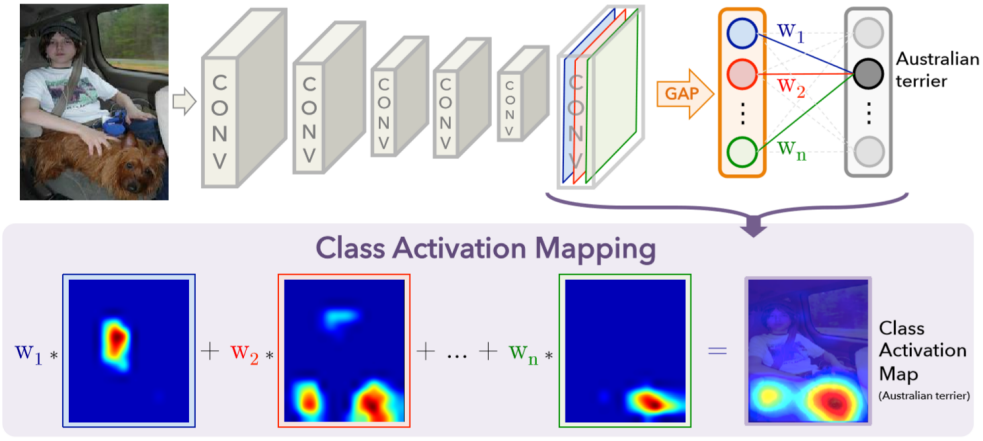
get attention-based model instantly by tweaking your own CNN a little bit more https://github.com/metalbubble/CAM/blob/master/pytorch_CAM.py

通过图片标签就能让卷积神经网络具有卓越的定位能力

It highlights the most informative image regions relevant to the predicted class.

CAM：预测的类别分数被映射回先前的卷积层以生成类别激活图（class activation maps），CAM凸显了特定类的区别性区域

CAM简单说就是不同空间区域的线性加权可视化。将类激活图的大小改变成输入图片的大小，就能清楚地看出与特定类最相关的区域。

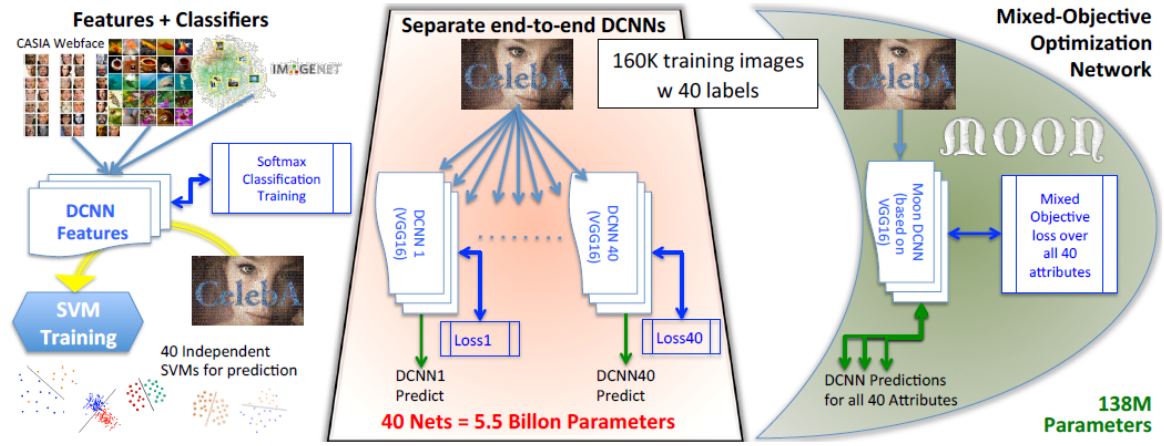


全局平均池化层（GAP）VS全局最大池化层（GMP）：高亮区域的不同。比起GMP的鼓励网络只关注1个discriminative part，GAP更鼓励网络识别对象的整个范围。因为当求平均数时，这个值可以通过找所有discriminative part来最大化激活值而低激活的part降低了其值。另一方面，对于GMP，所有图的低分区域（除了最有区分力的一个），都不会对得分有影响，因为你只取了max。GMP的分类性能与GAP相当，GAP的定位能力要强于GMP。

10. MOON : A Mixed Objective Optimization Network for the Recognition of Facial Attributes

ECCV 2016

University of Colorado at Colorado Springs



measure the difference between the source and target distributions 考虑两种集合分布的差异，而构造数据集时尽量使两者一致。

- Si: the **relative number** of occurrences of the positive Si+ and negative samples Si- in the training set
- Ti

$$p(i|+1) = \begin{cases} 1 & \text{if } T_i^+ > S_i^+ \\ \frac{S_i^- T_i^+}{S_i^+ T_i^-} & \text{otherwise} \end{cases} \quad \text{and} \quad p(i|-1) = \begin{cases} 1 & \text{if } T_i^- > S_i^- \\ \frac{S_i^+ T_i^-}{S_i^- T_i^+} & \text{otherwise.} \end{cases} \quad (4)$$

We need a loss function which additionally mixes all attribute predictions and simultaneously infers latent correlations between attribute labels and image data.

where $f_i(x)$ is the network output for attribute i, and for which the output dimensionality is the number of attributes M. Across an N element training set X with labels Y this yields:

$$L(X,Y) = \sum_{j=1}^{N} \sum_{i=1}^{M} p(i|Y_{ji}) \, ||f_i(X_j) - Y_{ji}||^2. \quad (7)$$

MOON incorporates attribute correlations and can adapt the bias of the training dataset to a target distribution.

11. Heterogeneous face attribute estimation: A deep multi-task learning approach

TPAMI 2017

山世光

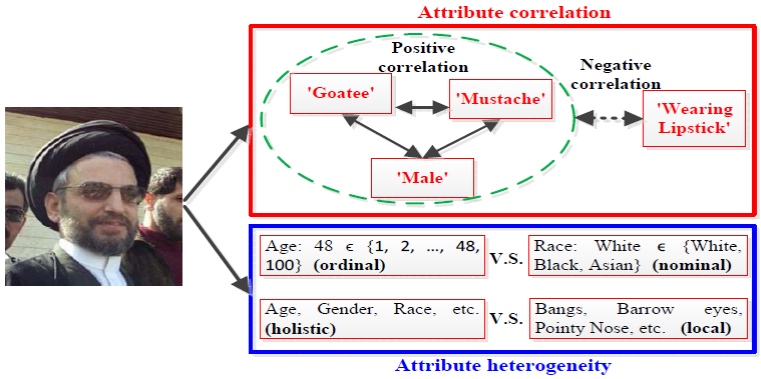


Fig. 1. Individual face attributes have both correlation and heterogeneity. While attribute correlation can be utilized to improve the robustness of attribute estimation, attribute heterogeneity should also be tackled by designing appropriate prediction models.

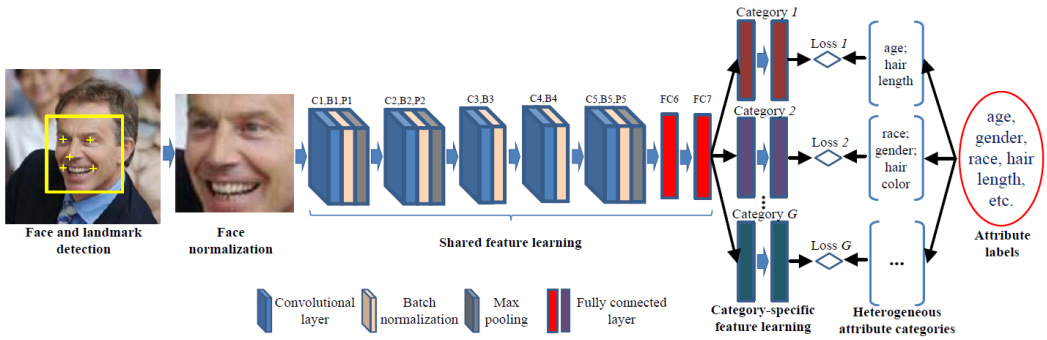


Fig. 2. Overview of the proposed deep multi-task learning (DMTL) network consisting of an early-stage shared feature learning for all the attributes, followed by category-specific feature learning for heterogeneous attribute categories. We use a modified AlexNet [11] with a batch normalization (BN) layer inserted after each Conv. layer for shared feature learning. The subnetworks are used to fine-tune the shared features towards the optimal estimation of individual heterogeneous attributes, *e.g.*, nominal vs. ordinal and holistic vs. local.

Network structure. For the shared feature learning network, we use a modified AlexNet network (5 convolutional (Conv.) layers, 5 pooling layers, 2 fully connected (FC) layers [11]) with a batch normalization (BN) layer inserted after each of the Conv. layers. Each of the category-specific feature learning networks contains two FC layers, and is connected to the last FC layer of the shared network.

没有回归距离，都是预测最终的标签属性

Binary Face Attributes

- one holistic nominal subnetwork
- seven local nominal subnetworks 依据local region来负责相关的属性集合

Though a number of commercial systems (e.g., Affectiva, Emotient, Face++, and Microsoft) provide estimates of attributes like age, gender and expression, the underlying algorithms used in commercial systems are proprietary; in addition, the databases used by these commercial engines are not (or no longer) available to the research community.

12. Fully-adaptive Feature Sharing in Multi-Task Networks with Applications in Person Attribute Classification

CVPR 2017

UCSD, IBM Research

多任务深度学习网络，一般是先设计网络有一些共享层，然后有多个分支学习不同的任务。论文从一个较瘦的网络开始，逐渐加粗。任务间进行选择性共享，挖掘那些任务之间更相关。thin网络使用SOMP初始化。

task-specific子网络或分支：浅层特征共享，深层特征task-specific，类似属性结构。计算量大，且受设计者主观认识影响。 automatically designs compact multi-task deep learning architectures with no need of discovering the possible multi-task architectures artificially.

13. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age

arXiv: 201908

UCLA

ResNet-34, ADAM lr=0.0001, dlib's CNN-based face detector

14. A Jointly Learned Deep Architecture for Facial Attribute Analysis and Face Detection in the Wild

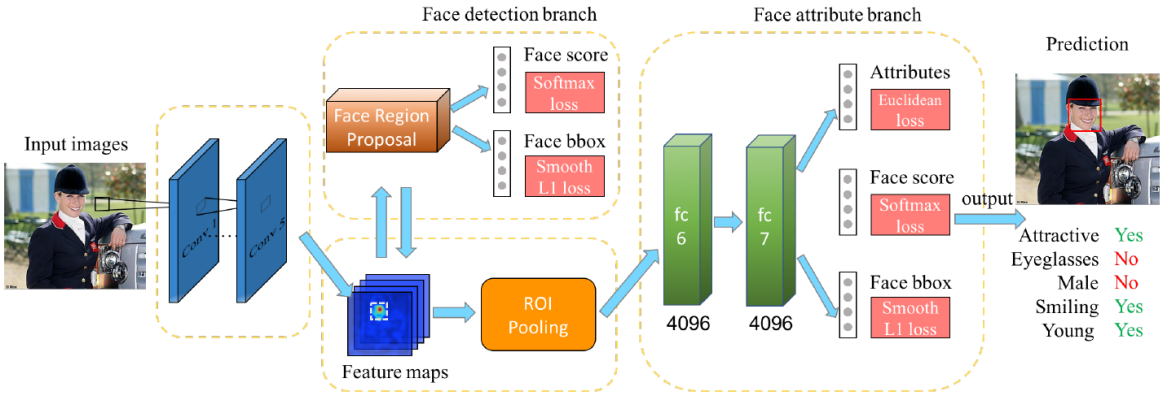
付彦伟，薛向阳

arXiv: 201707

with the preprocessing of face detector, the capability of facial attribute prediction has to heavily relies on the results of face detection.

idea: 施加face region proposal约束，使中间层的特征图具有检测人脸区域的响应。

- face detection branch
- facial attribute branch



By virtue of the face detection subnet, our architecture can directly predict the facial attributes from the whole images, rather than using the well-cropped images.

> Our network follows the art and design of VGG-16 [3]. Particularly, the kernel size, stride and the number of filters in convolutional layers (*conv1– conv5*) and the two fully connected layers (*fc6* and *fc7*) are exactly the same as the corresponding layers in VGG-16 architecture.

slide the face detection branch over the conv5 feature map

- face region proposal layer
- the RoI pooling layer

15. Face Attribute Prediction Using Off-the-Shelf CNN Features

arXiv: 1602

Sweden

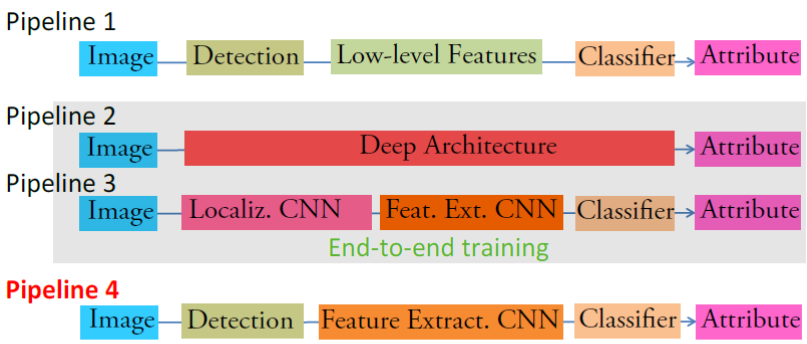


Figure 1: Potential pipelines of automatic attribute estimation.

predicting face attributes using CNNs trained for face recognition

并非所有的面部属性都适合只用high-level features from deep neural networks，毕竟有些是locally orientated

- conv
  - Google's FaceNet NN.1
  - VGG's very deep model http://www.robots.ox.ac.uk/~vgg/software/vgg_face/
- fc
  - 2 fc
  - dropout (p=0.5)

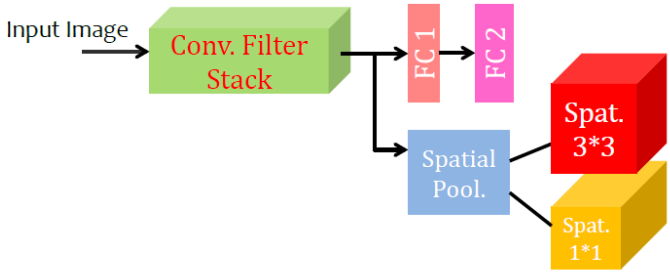


Figure 2: Pipeline of extracting deep representations from trained CNN. Intermediate features ($Spat.N \times N$, FC1) and final representation FC2 are extracted from the trained network. N for the side of deep feature map after extra pooling step. In total, 4 types of representations will be studied for face attribute prediction.

The output of the last conv. filter stack was selected as the representative of the **intermediate representations** since it was shown to have **the most discriminability and spatial information** for recognition and image retrieval.

binary linear SVM classifiers were trained directly for all levels of representations (i.e. FCs and Spat:0s) to classify face attributes

the intermediate spatial representations predicted 5 attributes (“Bags Under Eyes” , “Blurry”, “Mouth S. Open”, “Pale Skin” and “Narrow Eyes”) much better than the last FC representations.

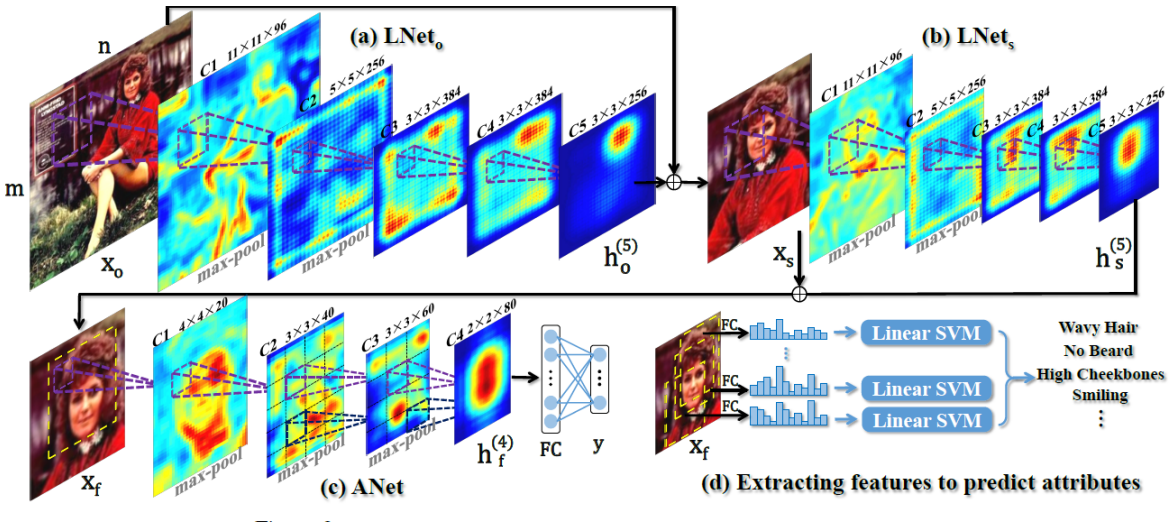
We believe the reason intermediate spatial representations outperformed on so many attributes is that these **human describable attributes** are more likely to be identified from the spatial information captured by human brains.

16. Deep Learning Face Attributes in the Wild

ICCV 2015

CUHK

不开源



- LNet: locates the entire face region in a coarse-to-fine manner with highly responded regions
    - trained in a weakly supervised manner with only image-level annotations
        - ho(5) indicates head-shoulders
        - hs(5) indicates faces
- ANet: extracts features for attribute recognition
    - pre-trained by face identity classification
    - SVM classifiers

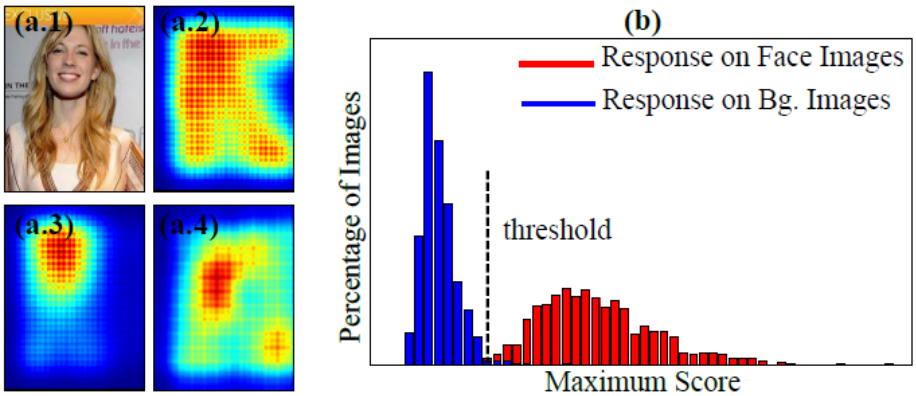


Figure 3. (a.1) Original image. (a.2)-(a.4) are averaged response maps in C5 of LNet$_o$ after pre-training (a.2), fine-tuning (a.3) and directly training from scratch with attribute tags but without pre-training (a.4). (b) Determine threshold.

Why rich attribute information enables accurate face localization?

When a subset of neurons are activated, they indicate the existence of face images with a particular attribute configuration. The neurons at different layers can form many activation patterns, implying that the whole set of face images can be divided into many subsets based on attribute configurations, and each activation pattern corresponds to one subset (e.g. ‘pointy nose’, ‘rosy cheek’, and ‘smiling’).

Therefore, it is not surprising that filters learned by attributes lead to effective representations for face localization.

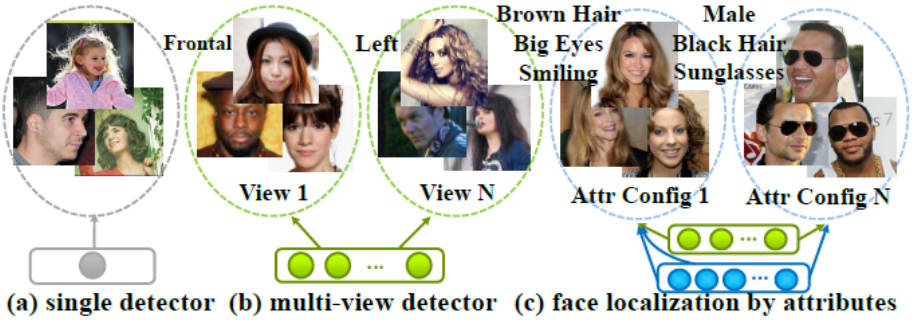


Figure 4. Face localization by attributes

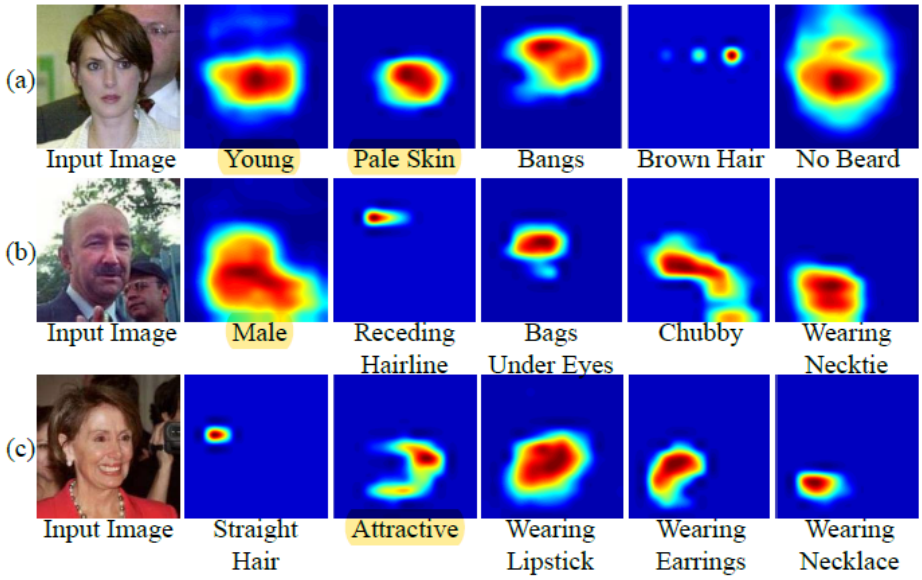


Figure 9. Attribute-specific regions discovery.

17. Characterizing the Variability in Face Recognition Accuracy Relative to Race

Florida Institute of Technology, University of Notre Dame

We report results from four face matchers: commercial SDKs “COTS-A” and “COTS-B”, and the popular CNN-based matchers VGG-16 and ResNet-50.

18. Slim-CNN: A Light-Weight CNN for Face Attribute Prediction

BMVC 2019

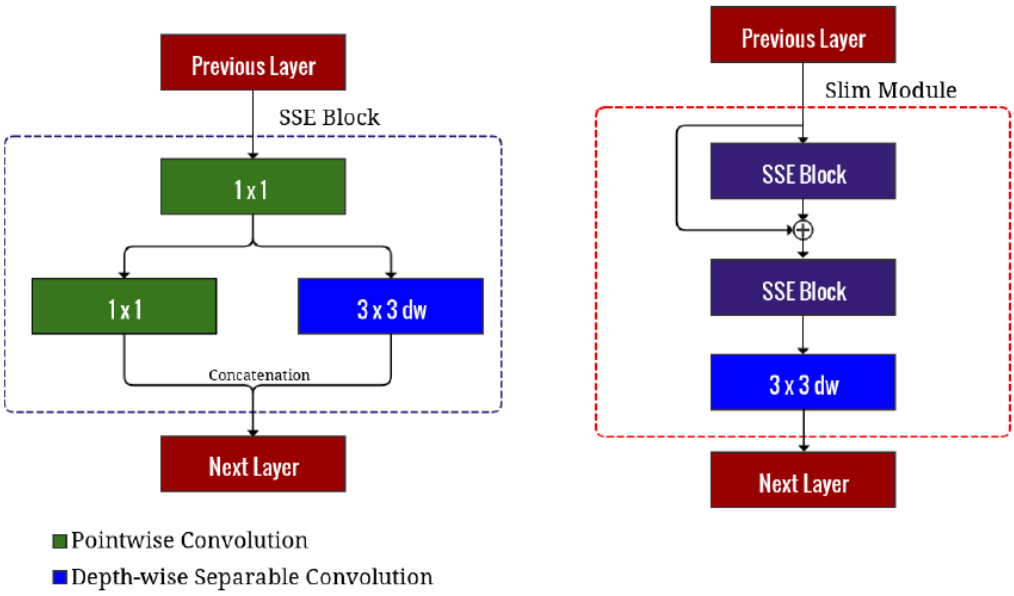
University of Central Florida

Related Work: Face Attribute Prediction 写的不错

paper整体排版很漂亮，写作很紧凑准确，没有废话，值得品味，像一件精美的艺术品。

Structural innovations:

- Multiple Branches
- Small Kernel：
- Skip-Connections



- SSE Block (Separable Squeeze-Expand)
- Slim Module
- Slim Network