Apparent Age Estimation from an Image

Group 5
Zhefu cheng
4591394
Sasank Kottapalli
4505718

Abstract

After decades of research, the real (biological) age estimation from a single face image reached maturity thanks to the availability of large public face databases and impressive accuracies achieved by recently proposed methods. The estimation of apparent age is a related task concerning the age perceived by human observers. Significant advances have been also made in this new research direction with the recent Looking At People challenges. In this paper we improved the performance of an existing model DEX[1] by (i) studying latest deep architectures and their performance regarding our age estimation model, (ii) using new dataset UTKFace [2] for training our model and (iii) Optimally using parameters and reducing training time by using depth-wise separable convolutions and global average pooling (iv) implementing a combination of Augmentation methods showing their significant contribution in training deep networks.

1. Introduction

Historically being one of the most challenging topics in facial analysis [3], automatic age estimation from face images has numerous practical applications such as demographic statistics collection, customer profiling, search optimization in large databases and assistance of biometrics systems. There are multiple reasons why automatic age estimation is a very challenging task. The most principal among them are an uncontrolled nature of the ageing process, a significant variance among faces in the same age range and a high dependency of ageing traits on a person. Recently, deep neural networks have significantly boosted many computer vision domains including unconstrained face recognition [4] and facial gender recognition [5]. However, the progress in unconstrained facial age estimation is much slower, due to the difficulty of collecting and labelling large datasets which is essential for training deep networks. The vast majority of existing age estimation studies deals with the problem of estimation of a persons biological age (i.e. objective age defined as the elapsed time since the persons birth date). However, in 2015 and 2016, ChaLearn Looking at People (LAP) conducted 2 versions of competition on apparent age estimation (i.e. subjective age estimated from a visual appearance of a person) v1 and v2 respectively. The organizers collected datasets of face images and developed a web service where people could annotate these images with an apparent age. More than 200 teams have participated in both the competitions together and the best approaches were based on deep Convolutional Neural Networks (CNNs). Our basic approach is taken from DEX model [1], winner of 2015 competition, but a tensorflow version of it. We improved the accuracy of this model by trying different architectures and finally using Xception model which uses 1/3 parameters compared to original model because of its depth-wise separable convolutions. Our model poses age as a classification task by using softmax expected value function and is able to obtain continuous values for age. Used UTKFace dataset to best fit this classifier. Explored many augmentation methods and improved overall performance of the model.

2. Related Work

2.1. Biological age estimation

The existing age estimation studies mainly focus on biological age estimation. The most used metric for evaluating systems of automatic estimation of a biological age is Mean Absolute Error (MAE), which is mean of the difference of predicted and actual age. Initially in early 2000s age was treated as classification problem and classical vision techniques were used. In 2007 [9], feature designing to describe the ageing pattern proved to be of particular importance. This was continued till 2011 trying different approaches. In 2014, [9] is one of the first works to apply CNNs for age estimation. Authors employed several shallow multiscale CNNs on different face regions. From then, there are many variances in using CNN, trying to improve the accuracy.

2.2. Apparent age estimation

Apparent age is age perceived by a human mind. Though it is highly co-related with biological age, they are different

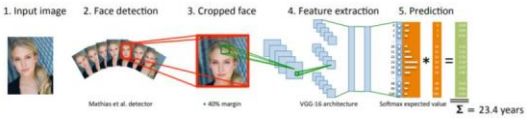


Figure 1: DEX

due to many factors. The first edition of the ChaLearn LAP AAE competition boosted the research in apparent age estimation by making public the first dataset with apparent age annotations of 4691 images. In the second edition of the competition, this dataset has been extended to 7591 images (4113 images for training, 1500 for validation and 1978 for test). Not only the number of images has increased, but also the age distribution has changed with respect to the first edition of the competition. The winning models of competition, DEX [1] in 2015 and OrangeLabs [6] in 2016, are designed according to the age distribution provided by the evaluation datasets. DEX [1] created a dataset named Imdb-Wiki with half a million images and used ensemble of CNN's where they finetuned their model on the evaluation dataset. OrangeLabs [6] in 2016, created a Children specialized deep CNN model and merged it with general one. As the evaluation dataset has high distribution of children images from ages 0-10, they successfully won the competition by a big margin.

3. Proposed Approach

3.1. Overview

We started with a tensorflow implementation of original DEX [1] model. This implementation was so basic that MAE was nearly twice that of original paper. Our approach was not to win any competition as it would lead us to develop our model one sided (depending on evaluation dataset), but to enhance the model understanding its basic structure so that its performance is consistent in any sort of evaluation. For this we used UTKFace dataset. Though the data is less, its distribution is more uniform from ages 0-100 suitable for our model's output layer. We experimented with different architectures from VGG-16 to latest DenseNet , we also tried reducing the parameters using depth-wise separable architectures and replacing fully connected layers. As we were using deep architectures to overcome the over-fitting we researched many state-of-art augmentation methods and generated an augmentation pipeline which is most sensible for an age estimation task. We drastically improved the performance of the model we initially used.

3.2. Original model and Our chosen model

Original model DEX [1] (Figure 1): Input image is passed through of-the-shelf face detector of Mathias et al [7] for face detection. To detect the best aligned face, face detector is run on all rotated versions between 60 and 60 in 5 steps. Highest face score version is selected and cropped with 40 percent margin and this image is passed through VGG-16 architecture for feature extraction. They used softmax expected value as output layer. It contains 101 neurons for ages from 0 to 100 and the probability values produced in these neurons are multiplied with the corresponding age number. All these values are added and the resulting number is the age of the given input image as a continuous value. We chose a tensorflow implementa-tion of this model. Though the basic model is same, there were some differences such as (i) Size of the image used was 64x64 compared to 256x256 in the original. (ii) Archi-tecture used was wide residual networks (WRN) instead of VGG-16 in the original (WRN is better). (iii) Evaluation was done on APPA-REAL[12] (2016 LAP AAE challenge dataset) dataset rather than LAP (2015 LAP AAE challenge dataset) dataset. (iv) There was no -error evaluation. (v) Model was not finetuned on evaluation dataset while the original model was finetuned on evaluation dataset with 20 random splits of 90:10 of train to validation sets called en-semble of CNN's.

3.3. Architectures

DEX [1] used VGG-16 architecture. As there are many high performance architectures available compared to VGG-16, we chose best versions of each variant such as (i) RESNET-50 (ii) Wide residual networks (WRN) (iii) Incep-tionV3 (iv) InceptionResNetV2 (v) DenseNet-201. We used Adam and SGD optimizers. We did a lot of hyperparameter tuning to get the best result out of each network.

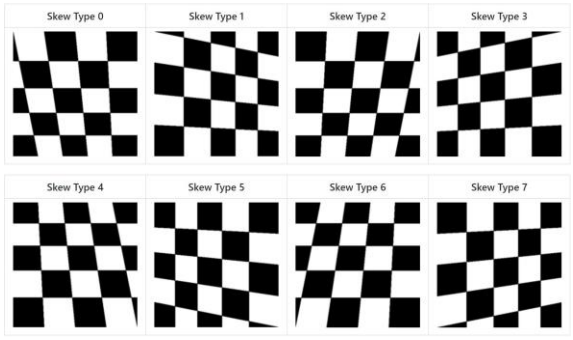
3.4. Parameter Reduction

Inspired by a emotion classification model [8] we decided to reduce our parameters by using depth-wise separable convolutions instead of standard convolutions and also replacing the fully connected layers with global average pooling. For this purpose we selected the Xception [9] architecture. Xception network efficiently uses the parameters. Though the network is also deep it managed to have less parameters and better performance compared to InceptionV3. We made Xception our primary architecture and optimized its performance using hyperparameter tuning. Xception takes 1.5x - 2.5x less time compared to other architectures used in our training.

3.5. Augmentation

To make our model invariant to a variety of conditions, such as different orientation, scale, brightness, etc, and also the deep networks we are using contains more parameters compared to the number of training samples, this leads to over-fitting and weakening its generalization ability. Using different augmentation methods gives us more data which

(a) Random erasing



(b) Random skew

Figure 2: Data augmentation

```
p = Augmentor.Pipeline()
p.flip_left_right(probability=0.5)
p.rotate(probability=0.5, max_left_rotation=5, max_right_rotation=5)
p.zoom_random(probability=0.5, percentage_area=0.95)
p.random_distortion(probability=0.5, grid_width=2, grid_height=2, magnitude=8)
p.skew(probability=0.5, magnitude=0.1)
p.random_color(probability=0.5, min_factor=0.8, max_factor=1.2)
p.random_contrast(probability=0.5, min_factor=0.8, max_factor=1.2)
p.random_brightness(probability=0.5, min_factor=0.8, max_factor=1.2)
p.random_erasing(probability=0.5, rectangle_area=0.20)
```

Figure 3: Data augmentation pipeline

appears to be unique and also settles the uncertainties in the testing dataset. Age labels of training images are also augmented as age itself is uncertain and this adds noise in the age labels.

Mixup [18] is a type of image augmentation methods that augments training data by mixing-up both of training images and labels by linear interpolation with weight , where is drawn from the Beta distribution.

$$X = X_1 + (1 \quad)X_2$$

$$y = y_1 \quad + (1)y_2$$

Random Erasing [19] is another type of image augmentation methods that tries to regularize models using training images that are randomly masked with random values. Figure 2(a) illustrates how random erasing works.

Other image augmentation techniques being applied are horizontal flip, size preserving rotation, random zoom, random distortion, random skew (Figure 2(b)), and random change of image color, contrast, and brightness. Figure 3 shows how a combined augmentation pipeline looks.

Labels are augmented by adding Gaussian noise to age labels. numpy.random.randn generates random values from the Gaussian distribution.

age = age + math:f loor(np:random:randn() 2 + 0:5)

4. Experiments

4.1. Dataset

4.1.1 UTKFace

UTKFace [2] is a large-scale face dataset with long age span (range from 0 to 116 years old) and more uniform distribu-tion of ages. It consists of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc. We use the UTKFace in-the-wild faces dataset, rather than the UTKFace tightly cropped faces dataset, to create face images that are cropped with a 0.4 margin around face regions. 23252 cropped images are split into 20000 for training and 3252 for validation.

4.1.2 APPA-REAL

APPA-REAL dataset [9] is from ChaLearn Looking at People (LAP) that held two apparent age estimation challenges in 2015 and 2016, and was used for the 2016 challenge. It contains 7591 images with associated real and apparent age labels, where apparent ages are based on around 250000 votes. The images, which contain cropped and rotated faces with a 0.4 margin, are split into 4113 training, 1500 vali-dation and 1978 test images. The test images are used for evaluating model performance. This dataset is made public in 2017 [9].

4.2. Experiment Details

Experiments are done on the UCF Newton HPC, with tensorflow-gpu environment of Anaconda2-5.3.0, and two NVIDIA Tesla V100-PCIE-16GB GPUs.

The CNN models are pre-trained on ImageNet, and have global average pooling enabled. For training, learning rate and batch size stay unchanged as 0.001 and 32, while num-ber of epochs is at least 40, depending on how fast the model converges. As InceptionResNetV2 has shown better results in early experiments, most of later experiments are done with it, such as setting SGD as optimizer instead of Adam, disabling data or label augmentation, testing out different data augmentation configurations in training. When the models are trained on only UTKFace dataset, each epoch takes around 7 minutes in average, but it increases to 11 minutes when trained on both UTKFace and APPA-REAL datasets. In most cases, a model requires over 6 hours to train, and training InceptionResNetV2 takes the most time due to its over 50 million trainable weights. Later, we made Xception as our main model and continued experimenting with it until we got to a convergence point in accuracy.

Models	Training Conditions	MAE Apparent	MAE Real	-error
VGG-16	Best Configuration	4.7859	6.0664	0.4187
ResNet50	Best Configuration	4.4327	5.8871	0.4085
WideResNet	Best Configuration	4.3637	5.9888	0.4103
DenseNet201	Best Configuration	4.1731	5.8891	0.4159
InceptionV3	Best Configuration	3.9467	5.4844	0.3823
InceptionResNetV2	Best Configuration	3.8119	5.4912	0.3747
Xception	Enhanced data aug (UTKFace)	3.8018	5.4501	0.3616 (best)
Xception	Enhanced data aug (UTK + APPA)	3.7997 (best)	5.3066 (best)	0.3709

Table 1: Evaluation results on test set of APPA-REAL in Best Configuration = Enhanced data aug (UTK + APPA)

Models	Training Conditions	MAE Apparent	MAE Real	-error
Xception	No augmentation	4.5092	5.8883	0.4191
Xception	No data augmentation	4.2037	5.7433	0.4029
Xception	SGD optimizer with best conf	4.0799	5.4264	0.3974
Xception	No label augmentation	3.978	5.6255	0.3722
Xception	UTKFace + APPA-REAL	3.8632	5.4091	0.3767
Xception	Enhanced data aug (UTKFace)	3.8018	5.4501	0.3616 (best)
Xception	Enhanced data aug (UTK + APPA)	3.7997 (best)	5.3066 (best)	0.3709

Table 2: Performance on different configurations

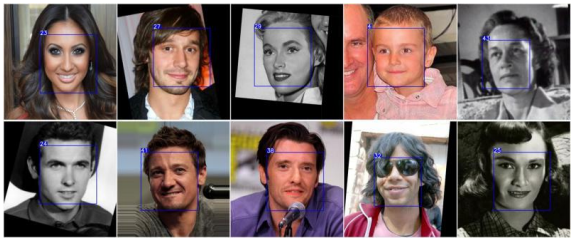


Figure 4: Sample age estimations on test set of APPA-REAL

Estimation	23	27	29	4	43	24	41	38	32	25
Apparent age	26.8	28.7	27.5	4.2	48.8	26.6	40.3	36.1	31.4	27.4
Real age	24	28	26	4	44	27	43	41	33	22

Table 3: Estimated, apparent and real ages for images in Figure 4. MAE apparent = 2.12, MAE real = 1.8

4.3. Evaluation Metrics

MAE. The mean absolute error (MAE) is the average of absolute errors between the estimated age and the labeled age.

-error. APPA-REAL dataset images are annotated with the mean and standard deviation of age votes by multiple voters. A normal distribution is fit with and of the votes for each image: [insert an equation]

For each age estimation, can be maximum 1 (the worst) and minimum 0 (the best). For a set of images, -error is computed in average.

4.4. Evaluation Results

Trained models are evaluated on test set of APPA-REAL, results are summarized in Tabel 2. Apparently, Inception and Xception models significantly outperform VGG-16, ResNet, WRN and DenseNet. InceptionResNetV2 with

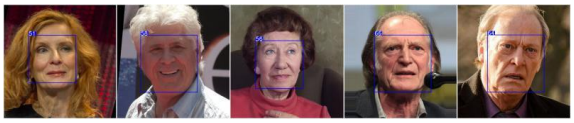


Figure 5: Sample age estimations of old people on test set of APPA-REAL

Estimation	51	58	56	66	64
Apparent age	61.8	64.9	71.3	70.8	66.4
Real age	61	62	73	72	64

Table 4: Estimated, apparent and real ages for images in Figure 5. MAE apparent = 8.04, MAE real = 7.4

residual connections shows better performance than InceptionV3 without such connections, 3.8632 to 3.9467 regard-ing apparent age MAE after being trained under the same condition. Xception model even yields slightly better evaluation results than InceptionResNetV2, benefiting from its depthwise separable convolutions, despite its parameters being nearly 1/3 of that of InceptionResNetV2. Adam optimizer is chosen as the default optimizer in training, and changing to SGD causes a 0.21 increase in apparent age MAE for Xception. The results also demonstrate the neces-sity of augmentation methods in training. After Xception being trained without any augmentation, it yields a 4.5092 apparent age MAE on test set of APPA-REAL, an over 0.6 decrease in performance. Moreover, data augmentation ap-pears to have a larger impact on model performance than label augmentation, as training with no data augmentation leads to a bigger decline in evaluation results than training without label augmentation, 0.3 worse regarding apparent age MAE. There is still more potential in augmentation, as an enhanced configuration of data augmentation methods leads to further improvement in age estimation. Overall, the smallest apparent age MAE 3.7997, smallest real age MAE 5.3066 and smallest -error 0.3616 all are achieved by Xception as the core architecture training with enhanced data augmentation.

Qualitative results are shown in Figure 4 and Figure 5. The sample age estimations are made by our best model with Xception as the core CNN architecture. Age estimations for images in Figure 4 have an apparent age MAE of 2.12 and a real age MAE of 1.8, which overall are good estimations. Among the samples, it is worth noting that our model estimates the age of a few old black-and-white images and noticeably an image of a person wearing sunglasses by close margins. It also accurately tells the age of a 4-year-old kid. However, we have noticed that our model tends not to perform so well on images of old people, as shown by the big MAEs for Figure 5. This might be due to the fact that despite the UTKFace dataset (taking the major portion in our training) has a more uniform age distri-

Methods	MAE
Sighthound [10]	5.76
OURS	6.45
Rothe et al. [1]	7.34
Microsoft. [13]	7.62
Kairos [14]	10.57
Face++ [15]	11.04

Table 5: Evaluation on Group[20] dataset and Adience[21] dataset

Position	Methods	-error	Finetuning on APPA-REAL
1	OrangeLabs[6]	0.2411	YES
2	palm-seu[16]	0.3214	YES
3	cmp+ETH[17]	0.3361	YES
4	WYU-CVL	0.3405	-
5	OURS	0.3616	NO
6	ITU-SiMiT	0.3668	-
7	Bogazici	0.3740	-
8	MIPAL-SNU	0.4569	-
9	DeepAge	0.4573	-

Table 6: Results of our method and other methods that participated in the ChaLearn LAP 2016 apparent age (AAE) challenge

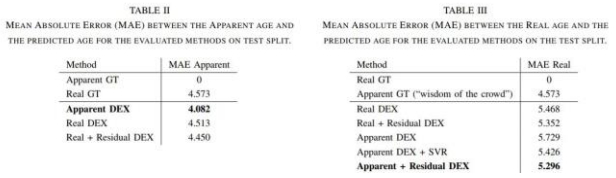


TABLE II
MEAN ABSOLUTE ERROR (MAE) BETWEEN THE APPARENT AGE AND THE PREDICTED AGE FOR THE EVALUATED METHODS ON TEST SPLIT.

Method	MAE Apparent
Apparent GT	0
Real GT	4.573
Apparent DEX	**4.082**
Real DEX	4.513
Real + Residual DEX	4.450

TABLE III
MEAN ABSOLUTE ERROR (MAE) BETWEEN THE REAL AGE AND THE PREDICTED AGE FOR THE EVALUATED METHODS ON THE TEST SPLIT.

Method	MAE Real
Real GT	0
Apparent GT ("wisdom of the crowd")	4.573
Real DEX	5.468
Real + Residual DEX	5.352
Apparent DEX	5.729
Apparent DEX + SVR	5.426
Apparent + Residual DEX	**5.296**

Figure 6: Results of Residual DEX (Updated version of DEX) model by same authors of DEX.

bution than many other datasets, the amount of images it has for people over 60 years old is still pretty small, hence, our model is not enough trained regarding that. As Table 4 shows, it tends to estimate old people a lot younger.

5. Baseline Methods Comparison

As presented in Table 6, LAP 2016 AAE challenge results, -error of our model achieved 5th place out of 105 participants. The top 3 models are finetuned on evaluation dataset, while we did not. They designed their models according to the distribution of evaluation dataset which makes them restricted. Orange labs [6] created a children specialized deep CNN model and merged it with general one. As the evaluation dataset has a high distribution of children images from ages 0-10, they successfully won the competition by a big margin. We did not design our model for competition but to achieve an overall performance. As shown in Figure 6, APPA-REAL dataset was release in 2017 along with this paper [9]. 2/3 authors of DEX model are key authors of this paper. They modified DEX [1] model to Residual DEX and evaluated it on APPA-REAL. Re-garding MAE apparent, our updation of DEX outperformed theirs by a margin of 0.2 and MAE Real being almost same. Finally to test our model on different datasets, we evalu-ated our model on Group and Audience dataset. As results shown in Table 5, we successfully outperformed the original DEX model (Rothe et al. [1]) by a margin more than 0.8 in MAE Real. sighthound [10] used 4M training images com-pared to our 30000+. That would be our only drawback. Given these results, we can say that although our model would not top 2016 LAP AAE challenge, we managed to achieve good overall performance.

6. Discussion

Besides the already mentioned CNN architectures, we have additionally tried a 2017 architecture called NASNet [11], which is a very deep architecture with over 88 mil-lion parameters. However, this expensive architecture de-manded an over large amount of memory to train, which exceeded the amount of memory available to us on the clus-ter, so eventually we did not successfully test it. Regarding augmentation methods, initially we tried changing image contrast by histogram equalization, which adjusts the con-trast of an image by changing the intensity distribution of image histogram to uniform distribution. But it turned out to be far less effective than changing the contrast randomly. We also tried shearing the training images, but shear was proved to be counterproductive after skew was already ap-plied. We added Gaussian noise to age labels, which was simple but effective. Later we tried multiple more complex configurations for the noise, but none surpassed the perfor-mance yielded through the initial Gaussian noise. We then had to accept the fact that sometimes simpler is better. How-ever, we believe there is certainly still room for enhancing the configuration of augmentation methods applied in this work, which may require thinking from a new direction. That could be part of the future improvement. One more aspect worth noting is that the size of our training images is less than 40 thousand, which is not big. A much larger dataset can likely enhance the model performance and lead to even better age estimation results.

7. Conclusion and Future Work

In this work, we presented modified version of DEX and successfully outperformed the original, outside LAP AAE challenge not only by using less number of parameters but also using a better distributed dataset which has signifi-cantly less images. We moreover outperformed an updated version of DEX named Residual DEX [9]. We started our work with MAE of 6.46 and significantly improved its per-

formance to MAE of 3.79 by (i) Using better architecture, (ii)better dataset, (iii) reducing parameters to 1/3 of original and (iv) By using pipeline of augmentation methods which showed significant improvement in results. In the future, we would like to use state of art face detection and alignment methods, facial landmarks, huge datasets manually adjusting their distribution uniformity. This would help our model in achieving even better results.

8. Acknowledgement

We have also made a Github repository for all code files and data, the link is: https://github.com/SashankMLAI/AgeEstimation

9. References

[1]R. Rothe, R. Timofte, and L. V. Gool, "DEX: Deep EXpectation of apparent age from a single image," in Proc. of ICCV, 2015.

[2]UTKFace-Dataset - http://aicip.eecs.utk.edu/wiki/UTKFace

[3]H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In Proceedings of IEEE International Conference on Biometrics, 2013.

[4]B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In Proceedings of European Conference on Computer Vision, 2014.

[5]G. Antipov, S.-A. Berrani, and J.-L. Dugelay. Minimalistic cnn-based ensemble model for gender prediction from face images. Pattern Recognition Letters, 70:5965, 2016.

[6]Antipov, Grigory, e.a.: Apparent age estimation from face images combining general and children-specialized deep learning models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2016)

[7]M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In ECCV. 2014.

[8] arXiv:1710.075

[9]E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, and R. Rothe, "Apparent and real age estimation in still images with deep residual regressors on APPA-REAL database," in Proc. of FG, 2017

[10] arXiv:1702.04280

[11] https://keras.io/applications/nasnet

[12] http://chalearnlap.cvc.uab.es/dataset/26/description/

[13] Microsoft-Face-API(https://www.microsoft.com/cognitive-services/en-us/faceapi.)

[14]Kairos (https://www.kairos.com/kairos-2.0/demos)

[15]Face++ (http://old.faceplusplus.com/demo-detect/)

[16]Huo, Z., Yang, X., Xing, C., Zhou, Y., Hou, P., Lv, J., Geng, X.: Deep age distribution learning for apparent age estimation.. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2016)

[17] Ui, M., Timofte, R., Rothe, R., Matas, J., Gool, L.V.: Structured output SVM prediction of apparent age, gender and smile from deep features. In: Proceedings of IEEE conference on Computer Vision and Pattern Recogni-tion Workshops, Las Vegas, USA (2016)

[18]H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in arXiv:1710.09412, 2017.

[19] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," in arXiv:1708.04896, 2017.

[20]Gallagher, A.C., Chen., T.: Understanding images of groups of people. In: CVPR. (2009)

[21]Eidinger, E., Enbar, R., Hassner., T.: Age and gender estimation of unfiltered faces. In: IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECU-RITY. (2013)