2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Uniface: A Unified Network for Face Detection and Recognition

Zhouyingcheng Liao[1], Peng Zhou[1], Qinlong Wu[2] and Bingbing Ni[1]
[1]Shanghai Jiao Tong University, Shanghai 200240, China
[2]China Mobile Suzhou Software Technology Co., Ltd, Suzhou 215000, China
patrickliao2007@gmail.com, zhoupengcv@sjtu.edu.cn, wotterlong@gmail.com, nibingbing@sjtu.edu.cn

***Abstract*—Typically, cropped and aligned face images are required as the input of a face recognition model. In contrast, popular object detectors based on deep convolutional network usually locate and classify objects simultaneously, which eliminates redundant computation. This work presents a single-network model called Uniface network for simultaneous face detection, landmark localization and recognition. We develop a feature sharing infrastructure for seamlessly integrate both the detection/localization module and the recognition module. To facilitate large-scale end-to-end training, we propose a method by encouraging top-level features of our model to mimic those of a well-trained single-task face recognition model. Comprehensive experiments on face detection, landmark localization and verification tasks demonstrate that the proposed network achieves competing performance in both face recognition benchmark (99.0% on LFW for a single model) and face detection benchmark (86.4% against 2000 false positives on FDDB for a single model).**

I. Introduction

Face recognition is one of the most active research topics in the area of computer vision. With the development of deep learning, a series of Convolutional Neural Network (CNN) based face recognition models have obtained great accuracy advances on LFW [1]. Current state-of-the-art methods, *e.g.*, DeepID [2], Facenet [3], DeepID2+ [4], basically follow a pipeline: face detection, face alignment and feeding the cropped and aligned face regions into a deep CNN for face recognition. Face detection models and face recognition models, both of which are trained to extract faces' features for analysis. Separating them into two models could be computationally wasting and inefficient. Many general object detectors [5] [6] solve the computation problem by simultaneous object detection and classification (Object detection refers to such simultaneous object detection and classification later in this paper). Furthermore, works in various fields of machine learning [7] [8] have shown that due to their inherent correlation, training similar tasks simultaneously could boost up all their performance.

Motivated by the above reasons, we propose a face recognition model which does not need cropped and aligned face images as input and is capable of simultaneous face detection and recognition. It should be noted that our work does not aim to use one model to achieve state-of-the-art performance in both face detection and face recognition, because in face detection benchmarks, *e.g.* FDDB [9], WIDER face [10], there are many extremely low-pixel and blurred faces whose identification could hardly be recognized. Excessively pursuing detection accuracy on this kind of faces could lead to an overfitting of our model.

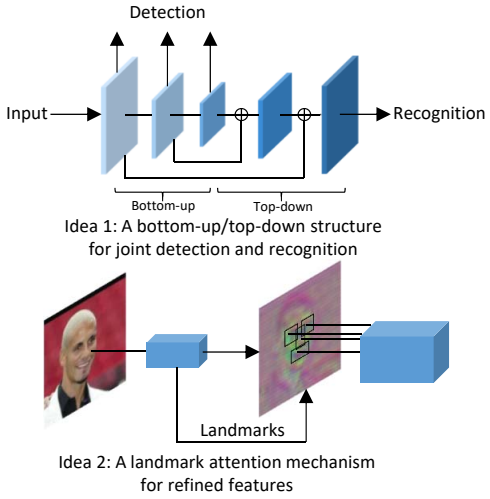

Fig. 1. **Two main ideas we present**. Above: we adopt a bottom-up/top-down architecture to extract sharing features. Below: we propose the landmark attention to extract more refined features for face recognition.

One of the most notable difference between object detection and face detection&recognition is that face recognition is much more fine-grained. Using existing object detectors in face detection&recognition directly could lead to a low accuracy in face recognition. To tackle this problem, we introduce a **bottom-up/top-down architecture** [11] and a **landmark attention mechanism** demonstrated in Fig 1 which makes it possible to train a joint face detection, landmark localization and face recognition network with competitive performance. Our model works as follow, a face detector embedded in our network will propose bounding boxes and landmark coordinates of detected faces. The predicted landmarks and bounding boxes act as an attention to tell the network where the key features lie in. With the landmark attention, the network will further extract more refined features and propose a 128-D embedding for face verification and identification.

Since few works that train face detection and face recognition simultaneously have been conducted before, hardly are there large-scale datasets with annotations of both face loca-

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

tions and face identifications. While face locations could be generated by the face detection algorithm, face identification could hardly be annotated. We adopt a mimic method to directly learn top-level features produced by a state-of-the-art single-task model. This makes ground-truth face identification annotations unnecessary and enables large-scale end-to-end training for simultaneous face detection and recognition.

II. Related Work

1) Face Recognition.: With the development of deep learning, multifarious models based on CNN have made a remarkable accuracy on LFW [1]. Face recognition is one kind of distance metric learning whose basic idea is to draw features of the same identity closer and features of different identities farther. The main differences of these face recognition models lie in the choices of the loss function. Facenet [3] presents triplet loss and hard negative exemplars mining. Softmax loss is combined into face recognition in [12], [4]. To learn more discriminative features, center loss [13] is proposed. Different from above ways, we directly set top-level features of a well-trained model as supervision and regress features of our model to it.

2) Face Detection.: In the field of general object detection, state-of-the-arts methods could be roughly divided into two categories, two-stage method [14] [5] and one-stage method [15] [6]. Two-stage methods, which coarsely propose a set of candidate boxes and then using a classifier to filter out the foreground boxes precisely, achieve high accuracy but are usually time-consuming. In contrast, One-stage methods, which directly propose the foreground boxes, are usually faster and simpler. Compared to the general object detector. Face detector, *e.g.* [16] [17] [18] [19], usually have shallower features because compared to the general object, faces have similar shapes and looks.

3) Multi-task Learning.: Since face detection, landmark localization and face recognition are done simultaneously, our model could be classified as multi-task learning [20]. Multi-task learning has been widely applied to many fields of machine learning. *e.g.* [21] [22] [23] [19]. These works demonstrate that inherent correlation between similar tasks could boost up each task's performance.

In recent years, various multi-task face-related models have been proposed. Hyperface [8] combine face detection, landmarks localization, pose estimation and gender recognition. [24] presents a model for jointly face attribute analysis and face detection. However, compared to face recognition, face attribute analysis is much easier. All-in-one Face [25] presents a multi-purpose algorithm for simultaneous face detection, face alignment, pose estimation, gender recognition, smile detection, age estimation and face recognition. Yet it still needs face region as an input of the network and it takes an average of 3.5s to process an image which could be extremely time-consuming. Our model achieves face detection, landmark localization and face recognition simultaneously and takes only around 8ms to process an image.

III. Uniface Network

We propose a single, unified CNN model called Uniface network, which is capable of simultaneous face detection, landmark localization and face recognition. We present two novel ideas to make it possible to train a joint face detection, landmark localization and face recognition in one network:

- We adopt a bottom-up/top-down architecture [11] in the backbone network. It makes different feature layers shared properly between face detection and face recognition.
- We propose landmark attention, which makes the network focus on the key features of the face. With this, the network could extract more refined features for face recognition.

Our network architecture is illustrated in Fig 2. An image without cropping and alignment is fed into the backbone network for basic feature extraction. We adopt a bottom-up/top-down architecture to unify face detection and face recognition. With the detection results, landmark attention will be applied to extract refined features of the face, which will be processed to propose a 128-D embedding for face recognition.

A. Backbone Network.

We construct the backbone network to make features of different depths shared properly by different tasks so that we can unify face detection and face recognition into one single network. First we use a CNN model pretrained on ImageNet [26] to extract basic features from the raw image. In this work, we choose Inception-v3 [27]. We truncate Inception-v3 before classification layers and add some extra convolutional layers to form feature maps whose sizes decrease by 2.

B. Bottom-up/Top-down Architecture.

Since the faces we intend to recognize is of multi-scale, simply using an invariant feature map for recognition could lead to inaccuracy. We introduce the bottom-up/top-down architecture [11] and construct it upon the backbone network.

Bottom-up/top-down architecture is first introduced in [11] to get a high-fidelity mask. We utilize it to unify face detection and face recognition into a single network and enable multi-scale face recognition.

In this architecture, the bottom-up pathway will be used to predict face's bounding boxes and landmarks and the top-down pathway will be used to construct a feature map for face recognition. Since the bottom-up pathway is supervised by the detection task, different feature layers on the bottom-up pathway should obtain different scale's semantics. The top-down pathway fuses features from each layer of the bottom-up pathway to construct a feature map F_r containing semantics of different scales. With this feature map, the recognition module is capable of recognizing multi-scale faces. In such way, the detection and recognition can be unified properly into one single network.

The specific construction process of the bottom-up/top-down architecture could be recursively formulated as:

$$F_{td}^{i+1} = Upsampling(F_{td}^{i} + F_{bu}^{i}), i \in \{1, 2, 3, 4, 5, 6\}, \quad (1)$$

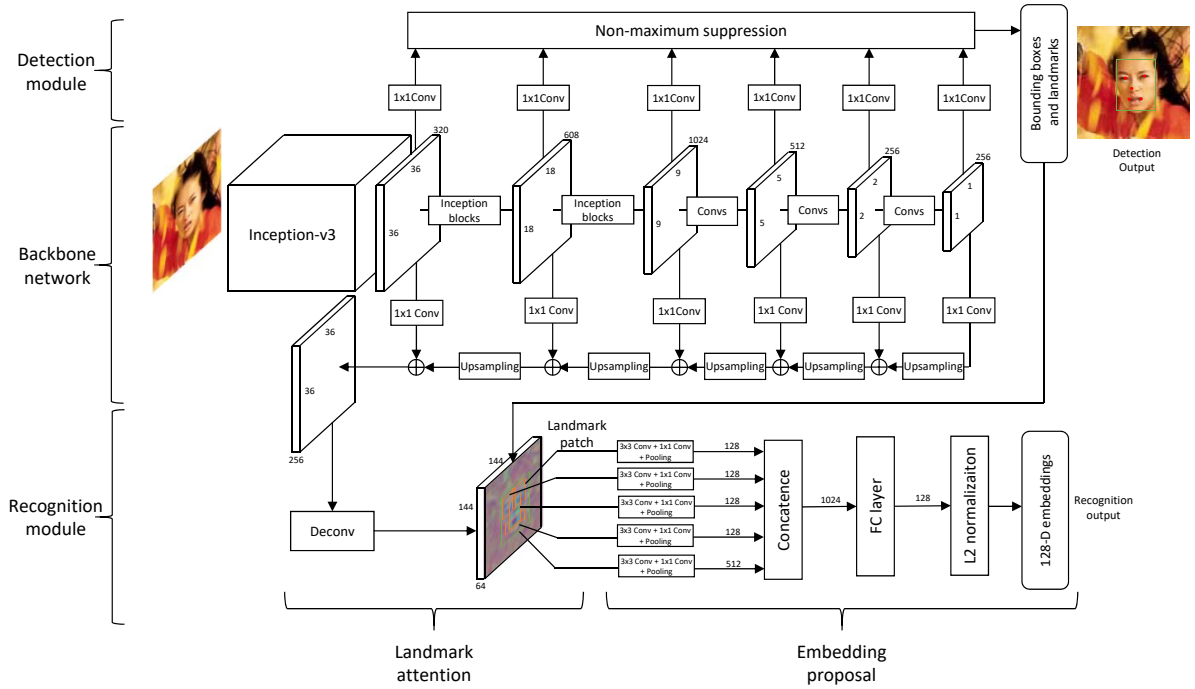

Fig. 2. **Network architecture.** An image without cropping and alignment is fed into our network. An Inception-v3 followed by a bottom-up/top-down architecture is used to extract sharing features. With the detection results, landmark attention will be used to extract refined features of the face, which will be processed to propose the 128-D embedding.

where F_{td}^i denotes the i-th feature layer of the top-down pathway, F_{bu}^i denotes the i-th feature layer of the bottom-up pathway and $Upsampling$ we use here is a nearest neighbor upsampling. The feature map for face recognition $F_r = F_{td}^6$. In the bottom-up pathway, sizes of feature maps decrease by a set of pooling layers with a stride of 2. Correspondingly in the top-down pathway, sizes of feature maps increase with a stride of 2. For a higher resolution, we apply two extra deconvolution layers on F_r to scale it to $64 \times 144 \times 144$. F_r constructed in this way will be of both high resolution and rich semantics.

C. Detection Module.

The detection module basically follows the paradigm of SSD [6]. For multi-scale face detection, we utilize the bottom-up pathway in the backbone network. A set of convolutional filters are applied on each layer of the bottom-up pathway to obtain the multi-scale detection results. The kernel size is $3 \times 3 \times p$ for a feature layer with p channels.

In SSD, each feature map cell is associated with a set of anchor boxes with different aspect ratios and sizes. Selecting aspect ratios and sizes reasonably is crucial to detection performance. Bounding boxes of faces have similar aspect ratios, which means there are fewer candidate aspect ratios. Via statistics on the training datasets, we find that aspects ratios of faces are mostly around 0.8. Therefore, we choose aspect ratios $a_r \in \{0.95, 0.8, 0.65\}$.

For each anchor box, there are $(2 + 4 + 10)$ outputs, which consist of 2 class(faces and backgrounds) scores, 4 bounding boxes offsets and 10 landmark localization offsets relative to the anchor box respectively. During inference, a non-maximum suppression will be applied to eliminate redundant boxes. Proposed detection results will then be used to extract refined features of faces.

D. Landmark Attention.

In our network, the goal of face recognition is to learn an embedding $y_i = f(F_r, d_i) \in R^{128}$, where F_r denotes the feature map for landmark attention and d_i denotes the i-th face's detection result. We propose a method called landmark attention for fine grained face feature extraction. Compared to ROI pooling [5], landmark attention utilizes face's landmarks to extract more refined features of the face.

After the feature map for recognition is constructed, we utilize the predicted bounding boxes and landmarks to extract refined features of faces. More specifically, for each detected face, we first exploit an ROI pooling on the face region in F_r to produce a pooled region with size $64 \times 48 \times 48$, which is followed by two conv-relu-pooling layers to produce a 512-D vector. Besides, the predicted landmarks will be used to get a fine grained attention. ROI pooling will be applied on the adjacent areas of each landmark to produce four $64 \times 14 \times 14$ regions. Similarly, after two conv-relu-pooling layers, four 128-D vectors will be made. These vectors will be concatenated to form a 1024-D vector, followed by a fully-connected layer to produce the final 128-D embedding.

E. Training

In our model, four outputs, class scores, bounding boxes offsets, landmark offsets and identification embeddings will

be made simultaneously. Accordingly, four kinds of loss will be made, which we denote by $L_{cls}, L_{bb}, L_{lm}, L_{id}$ respectively and we use hyperparameters λ_i to weight each kind of loss. Total loss is

$$L = \frac{1}{N}(\lambda_1 L_{cls} + \lambda_2 L_{bb} + \lambda_3 L_{lm} + L_{id}). \quad (2)$$

Next, we will first introduce losses in the detection module L_{cls}, L_{bb}, L_{lm}.

1) Detection Loss.: Our anchor boxes matching strategy is similar to SSD's, so they will not be detailed here. Briefly, we match each ground truth face box to the anchor boxes with the maximum Jaccard overlap which should be greater than a threshold. Assuming k anchor boxes matched, the left $(M-k)$ anchor boxes represent the background, where M denotes the total number of anchor boxes.

L_{cls} is the softmax loss between the background and the face. To avoid imbalance between positive and negative samples, we select the hardest, which have the highest confidence scores, $3k$ background boxes for calculating L_{cls}. L_{bb} is a Smooth ℓ_1 loss [14] of deviation of ground truth boxes relative to matched anchor boxes. Following this idea, L_{lm} is a Smooth ℓ_1 loss of deviation of landmark coordinates relative to matched anchor boxes.

2) Recognition Loss.: For an image with M faces, its embedding loss is

$$L_{id} = \frac{1}{M}\sum_{i=1}^{M} \|f(F_r, d_i) - g(x_i)\|_2^2, \quad (3)$$

where x_i denotes i-th face region cropped and aligned from the original image and g denotes a well-trained embedding for face recognition. This loss could be regarded as a mimic loss [28].

The mimic method has been widely used in model compression and acceleration. We apply it in multi-task learning for two reasons. Firstly, to the best of our knowledge, large-scale datasets for both face detection and recognition does not exist currently. With the mimic method, face identification annotations become unnecessary. Thus, we could train face detection and face recognition tasks end-to-end. Secondly, compared to sparate face detection and recognition models, our model is smaller and runs faster. In other words, our model could be seen as a compressed model in some way, which makes mimic method effective.

IV. Experiments

To evaluate the recognition ability of our model, we use LFW [1] as the benchmark of our model. To demonstrate that our model has the ability to detect faces, we also test it on FDDB [9]. Furthermore, we design three baseline models (Fig 3) to demonstrate the effectiveness of each key component in our model.

A. Experiment Settings

We use MS-Celeb(without alignment) and AFLW [29] as our training datasets. We randomly sample around one million images from MS-Celeb. Since we conduct mimic method to optimize the recognition task, ground truth of identity is not needed. Instead, we crop face regions of each image and feed the cropped area into a well-trained Facenet [3] model (with the accuracy of 99.4% of LFW). The proposed 128-D embeddings will be set as supervision of our model. Since MS-Celeb does not contain landmark annotations, we apply MTCNN [19] to get the landmark coordinates of each face.

Our model is implemented with MXNet [30]. All training and inference are carried out on a single Nvidia Titan X Pascal. For each input image, we pad it with 0 to a square and resize it to 300×300. Then, the RGB image is normalized to follow the standard normal distribution. The training of our model starts with a learning rate of 0.1. When the loss goes steady, we adjust the learning rate by dividing it by 10. While the batch size is 32, it takes around 120k iterations to finish the complete training.

TABLE I
Results on LFW.

Methods	Images	Aligned	Networks	Accuracy
DeepFace	4M	3D	4	97.35%
DeepID	203K	2D	60	97.45%
DeepID2	203K	2D	25	**99.15%**
DeepID2+	290K	2D	25	**99.47%**
Facenet	260M	**No**	1	98.87%
Facenet	260M	Yes	1	**99.63%**
Ours(Model A)	1M	**No**	1	**98.98%**
Ours(Model B)	1M	**No**	1	97.93%
Ours(Model C)	1M	**No**	1	97.50%
Ours(Model D)	1M	**No**	1	95.65%

B. Evaluation on Face Recognition

We use LFW as the benchmark of our models. LFW is a conventional face recognition benchmark, which contains 13233 images of 5749 identities. Following the protocol of unrestricted, labeled outside data, we use 10-fold cross validation to calculate the accuracy. For a complete test of our model, we do not use any external crop or alignment tools to process the images. The original images are directly fed into our network. All our models evaluated on LFW are trained on sampled MS-Celeb.

The primary model achieves the accuracy of $98.983\% \pm 0.389$, which is a competitive performance, especially for models without extra detectors. By contrast, Facenet [3] using the fixed center crop achieves the accuracy of 98.87%. Comparison of models tested on LFW is detailed in Table I.

1) A Robustness Evaluation.: Though our model achieves impressive performance on LFW, we could still not confirm that our model has strong generalization ability in various scenarios. Since faces in LFW are of similar sizes and positions, we design an experiment in which we randomly crop or pad images on LFW so that faces are distributed in various sizes and various positions. In detail, for each 250×250 image is LFW, we randomly crop or pad at most 60 pixels for each side. Then we carry out the 10-fold test on this processed dataset.

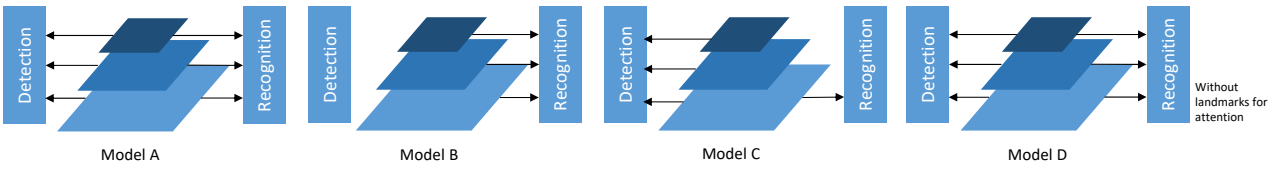

Fig. 3. **Comparison between primary model and baselines.** Model A: the primary model. Model B: we separate detection and recognition into two networks. Model C: Feature fusion is eliminated. Model D: landmark attention is eliminated.

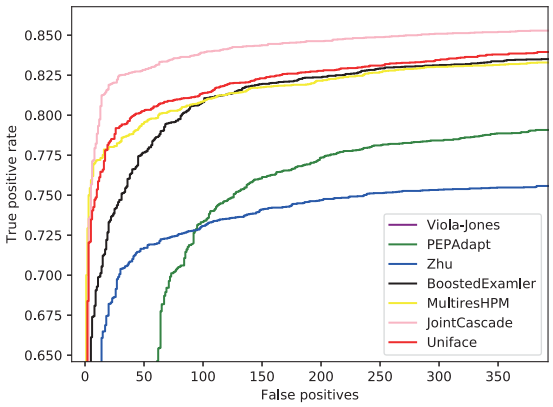

Fig. 4. **Results on FDDB.** The true positive rate of our model is 86.4% against 2000 false positives.

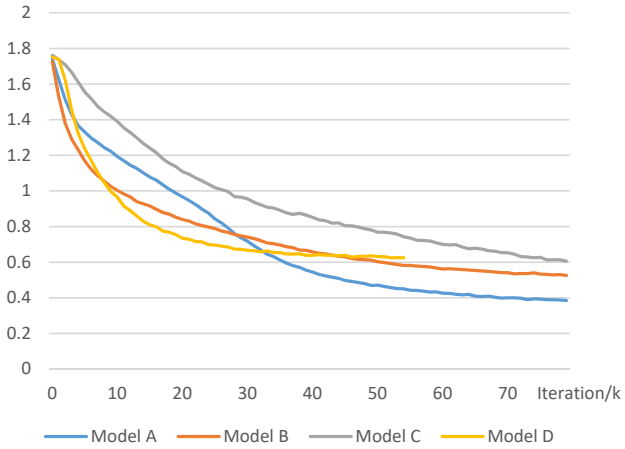

Fig. 5. **Identification losses of the primary model and three baseline models.** We draw losses of the primary model and three baseline models. It is obviously that the primary model has better convergence.

Our model achieves the result of $98.62\% \pm 0.56$. It is proved to be of great robustness.

C. Evaluation on Face Detection

To validate the detection performance of our model, we train our model using AFLW dataset and test it on FDDB [9]. FDDB measures the performance of models by computing an ROC curve by varying the threshold of face scores. We compare our model with some methods whose results are presented in FDDB result page. The result is illustrated in Fig 4.

D. Ablation Studies

To demonstrate the effectiveness of our model, we remove some key components from our primary model (Model A) to form the baseline models which are illustrated in Fig 3. To validate the performance boost brought by multi-task learning, we disassemble the detection and recognition part (Model B). We remove the top-down pathway to verify the effectiveness of feature fusion (Model C). We replace the landmark attention with conventional ROI pooling to verify the effectiveness of the landmark attention (Model D).

In Fig 5, we show the descent of recognition loss of the primary model and baseline models. It is evident that the primary model has better convergence.

1) The Effect of MTL.: To demonstrate the boost brought by multi-task learning, we will compare Model A with Model B. It is noted that we only design experiments to validate the boost brought by joint face detection and face recognition since joint detection and landmark localization has been proved effective by previous works like [19]. Model B has a similar structure with Model A, while we separate the detection and recognition part that they have no correlation anymore. For the training of Model, we deploy the parameters exactly the same as Model A. The final result of Model B is $97.93\% \pm 0.58$ on LFW. Its error rate is around twice as Model A's.

2) The Effect of Feature Fusion.: We adopt a bottom-up/top-down architecture to extract features and fuse features from different layers. This brings a huge advance in accuracy. For a fair comparison, we design a model without feature fusion (Model C). In Model C, we substitute the fused feature map with a feature map with the shape of $9 \times 9 \times 1024$. The rest parts of the network remain unchanged. Its final accuracy achieves $97.50\% \pm 0.62$.

3) The Effect of Landmark Attention.: In Model D We remove the landmark attention and only use the bounding boxes for ROI pooling. To make it fair, we double the number of channels of the rest recognition module to keep the parameter's number unchanged. Its final accuracy is $95.65\% \pm 0.85$, which is far behind the performance of Model A. It indicates the huge boost brought by landmark attention.

TABLE II
INFERENCE TIME OF DIFFERENT MODELS.

Model	Task	GPU	Speed/FPS
Faceness	Face Detection	Titan Black	20
MTCNN	Face detection and landmark localization	Titan Black	99
All-in-one	Face detection, attribute analysis and recognition	Titan X	0.286
Faster-RCNN	Object detection	Titan X	7
Ours	Face detection, landmark localization and face recognition	Titan X	**120**

E. Inference Time

In our model, low-level features are shared between face detection and recognition so the network only runs once for each image while in separate detection and recognition models, the recognition network runs N times for an image with N faces. Thus our model has great superiority in speed. We test the speed of our model with the batch size of 32 on a Titan X GPU. Comparison with other methods is listed in Table II.

V. CONCLUSIONS

In this paper, we present Uniface network for simultaneous face detection, landmark localization and recognition. It achieves the accuracy of 99.0% on LFW and reaches 120 FPS on a single GPU. We apply bottom-up/top-down architecture and landmark attention mechanism and validate their effectiveness in our task. We further validate that the inherent correlation between face detection and face recognition could bring a boost for both tasks. For the training of multi-task face detection and recognition, we set an example of adopting the mimic method to optimize the recognition task.

ACKNOWLEDGMENT

This research was supported by China Mobile Suzhou Software Technology Co., Ltd. We are immensely grateful to them for their assistance and guidance.

REFERENCES

[1] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[2] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[4] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2892–2900.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[7] R. Caruana, *Multitask Learning*. Boston, MA: Springer US, 1998, pp. 95–133.

[8] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *arXiv preprint arXiv:1603.01249*, 2016.

[9] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.

[10] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *European Conference on Computer Vision*. Springer, 2016, pp. 75–91.

[12] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[13] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[14] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[16] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[17] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[18] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5325–5334.

[19] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[20] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[21] Q. Zhou, G. Wang, K. Jia, and Q. Zhao, "Learning to share latent tasks for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2264–2271.

[22] Y. Yan, E. Ricci, R. Subramanian, G. Liu, and N. Sebe, "Multitask linear discriminant analysis for view invariant action recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5599–5611, 2014.

[23] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European Conference on Computer Vision*. Springer, 2014, pp. 94–108.

[24] K. He, Y. Fu, and X. Xue, "A jointly learned deep architecture for facial attribute analysis and face detection in the wild," *arXiv preprint arXiv:1707.08705*, 2017.

[25] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 17–24.

[26] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[28] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Advances in neural information processing systems*, 2014, pp. 2654–2662.

[29] P. M. R. Martin Koestinger, Paul Wohlhart and H. Bischof, "Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization," in *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[30] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.