Learning Discriminative and Complementary Patches for Face Recognition

Zhiwei Liu[1,2], Ming Tang[1,3], Guosheng Hu[4,5], Jinqiao Wang[1,3]

[1]National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China, [2]University of Chinese Academy of Sciences, Beijing, China, [3]Visionfinity Inc., ObjectEye Inc., Universal AI Inc., [4]AnyVision, [5]Queen's University Belfast,

***Abstract*— The ensemble of convolutional neural networks (CNNs) has widely been used in many computer vision tasks including face recognition. Many existing ensembles of face recognition CNNs apply a two-stage pipeline to target performance improvement [10], [20], [22], [23], [29]: (1) it trains multiple CNNs separately with many face patches covering different facial areas; (2) the features derived from different models are aggregated off-line by different fusion methods. The well-known face recognition work, DeepID2 [20] trains 200 networks based on 200 arbitrarily chosen facial areas and chooses the best 25 ones to achieve impressive performance. However, it is very time-consuming to train so many networks. In addition, a brute-force like way of choosing facial patches is used without knowing which face patches are complementary and discriminative. It might be lack of generalization capability for cross-database applications. To solve that, we propose a novel end-to-end CNN ensemble architecture which automatically learns the complementary and discriminative patches for face recognition. Specifically, we propose a novel Patch Generation Engine (PGE) with Patch Search Spatial Transformer Network (PS-STN) and ROI shrunk loss to perform the patch selection process. ROI shrunk loss enlarges the distance of learned features in spatial space and feature space and learn complementary features. In order to get final aggregated feature, we use a supervised fusion module named Two Stage Discriminative Fusion Module (TSDFM) which effective to capture the global and local information and further guide the PGE to learn better patches. Extensive experiments conducted on LFW and YTF datasets show the effectiveness of our novel end-to-end ensemble method.**

I. INTRODUCTION

Deep neural network ensemble is widely used for various computer vision tasks. The ensemble of networks can usually improve the performance greatly because different networks can capture complementary information. In the field of *object recognition*, network ensemble is widely used for many famous convolutional neural networks (CNNs) such as VGGNet [18], GoogleNet [24] and ResNet [3]. The fusion of the features derived from multiple networks can effectively improve the performance because the information extracted from these networks are complementary. For *fine-grained object recognition*, the bilinear model [9] achieves promising performance. This model [9] contains two CNNs which capture different parts of input images, and the features extracted by two CNNs are fused by outer product. For *action recognition*, the well known two-stream CNN [17] actually contains two complementary CNNs: one captures

978-1-7281-0089-0/19/$31.00 ©2019 IEEE

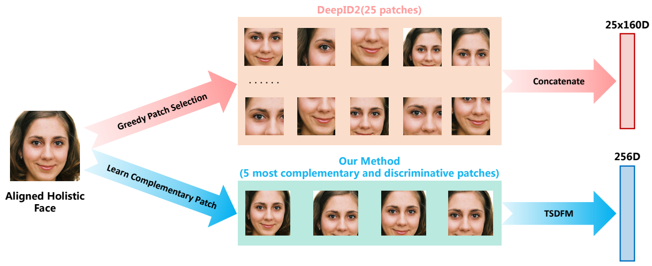

Fig. 1. The Difference between the traditional face recognition framework and our end-to-end ensemble framework

the appearance information of video frames and the other captures the motion information. This two-stream CNN is the first deep learning method which achieved comparable performance against the best handcrafted feature-based method [28]. Network ensemble is also applied to many other tasks, such as person re-identification [2], pose estimation [33] and etc.

Face recognition is a classical computer vision task, not surprisingly, network ensemble (or more general feature ensemble) is also successfully applied in this field. In the handcrafted feature era, the well-known multi-scale (MS) features, such as MS local binary patterns (LBP) [8] and MS local phase quantization (LPQ), are actually the fusion of features of different scales. MS-based features complement each other because they can capture texture information of different resolutions. In the deep learning era, the most representative work for network fusion is DeepID2 [20]. Motivated by the fact that different facial components (eyes, nose, etc) can provide complementary information, in [20], the training images of different networks are image patches covering different facial components. 200 CNNs are trained using 200 such patches, therefore, different CNNs can capture the information from different facial areas. Afterwards, 25 best performed networks out of 200 ones are chosen for feature extraction. DeepID2 achieves great success in terms of face recognition accuracy in LFW database [6]. Specifically, the accuracy of the worst and best networks are 86.63% and 96.33% on LFW respectively; while the fusion of 25 networks are 99.15%. Other deep metric learning algorithms with high performance like SphereFace [10] also make further promotion by ensembling different patch models. This fully proves the robustness of complementary information from different patches. However, it is hard to apply DeepID2 to real world because of the high computational complexity of 200, even 25 networks during both training and inference process. This high computational complexity results from our limited knowledge of which patches are the most complementary. The lack of insights into the way of achieving the most complementary patches leads to the brute-force like patch choosing process.

To solve the problem of DeepID2 and achieve promising network ensemble performance, in this work, we propose to automatically learn the complementary and discriminative patches to avoid the arbitrary patch choosing process, and further reduce the computational complexity in training and inference process, as shown in Fig. 1. It is observed that the image patches used for DeepID2 can actually be obtained from an aligned face image via translation, scale and cropping operations. Based on this observation, the patch selection process is learnable if we can learn the parameters of these spatial manipulations (translation, scale). To incorporate spatial manipulations into an end-to-end CNN training, we use constrained spatial transformer network (STN) [7], which was originally used for alignment-free object recognition and learn to transform discriminative regions for performance improvement. In this work, we proposed a novel CNN based framework with a Patch Generation Engine (PGE) followed by 5 face recognition sub-networks. The proposed PGE is composed of the Patch Search Spatial Transformer Network (PS-STN) and ROI shrunk loss. Benefited from the STN mechanism, we use Patch Search Spatial Transformer Network (PS-STN) to dynamic adjust patch regions and transform different discriminative face patches as the input of the following sub-networks. To learn the complementary features and avoid the learned patches converging to the same area, we introduce a ROI shrunk loss which enlarges the distance of learned features in spatial space and feature space.

The learned features from different face patches need to be fused properly, many existing methods like DeepID2 [20] and SphereFace [10] simply concatenates them to form a new feature. This off-line feature fusion method is not end-to-end and has no relevance to the face patches and the feature extraction networks, which is obviously not optimal in human's cognition. In this work, we use a supervised Two Stage Discriminative Fusion Module (TSDFM), which is effective to capture the global and local information to fuse different patches, as shown in Fig. 1. TSDFM also has ability to guide the proposed PGE searching better face patches which is demonstrated in our further experiments.

Our contributions can be summarized as:

- The existing face network ensemble method, such as DeepID2, uses a brute-force like way to select a large number of patches, In this work, we propose a novel deep framework which can automatically learn the complementary and discriminative patches for face recognition.
- In order to learn complementary and discriminative face patches, we propose a novel Patch Generation Engine(PGE) which has PS-STN to make the local facial patch selection process learnable and ROI shrunk loss to enlarges the distance of learned features in spatial space and feature space.
- Inspired by humans face recognition process, we propose to use a supervised Two Stage Discriminative Fusion Module (TSDFM) for feature fusion. TSDFM is end-to-end trained in our overall framework and guide PGE to search better face patches.
- Extensive experiments conducted on LFW database show the proposed architecture outperforms than traditional ensemble method by a large margin. Benefited by our effective end-to-end ensemble method, we also achieves comparable face recognition performance against state-of-the-art method on LFW database.

II. RELATED WORK

Recently, the introduction of deep learning models has greatly promoted the development of the face recognition technology. DeepFace [26] first demonstrated the effectiveness of data driven deep learning method and train a CNN model with locally connected layers to capture different local features. They also apply a novel 3D face alignment method to normalize the input faces under various postures. This technology is capable of handling out-of plane rotations and overcome some shortcomings of 2D alignment methods.

Another impressive work is DeepID series [20], [22], [23]. Unlike DeepFace whose features are learned by one single big CNN, DeepID first proposed to concatenate multiple feature which is extract by CNN models trained from various face regions. Both RGB and grey patches are used to extract DeepID features. The region of each patch are defined artificially and selected by a greedy way. The improvements of their results show that the features of patches are complementary and can be aggregated to produce a final discriminative feature. Inspired by the architectures of DeepFace and DeepID, many other face recognition methods apply deeper CNN models and provide additional boost to the performance such as [10], [23], [29].

Face verification is to decide whether two face images represent the same person or two different people. In view of this, many deep metric learning based algorithms take pairs of face images as input to learn a feature embedding where positive pairs are closer and negative pairs are far apart. DeepID2 [20] trains CNN models by using contrastive loss and softmax loss jointly which considers both identification and verification information. FaceNet introduces triplet loss to learn the metric using hard triplet face samples. CenterFace [28] proposes center loss to simultaneously learns a center for deep features of each class and penalizes the distances between the deep features and their corresponding class centers. In order to improve feature discrimination, large margin softmax (L-Softmax) [11] proposed to add angular constraints to each identity. Angular softmax (A-Softmax) [10] improves L-Softmax and achieves better performance on a series of open-set face recognition benchmarks by normalizing the weights. All the purposed of the methods above is to increase intra-class distance and decrease inter-class distance

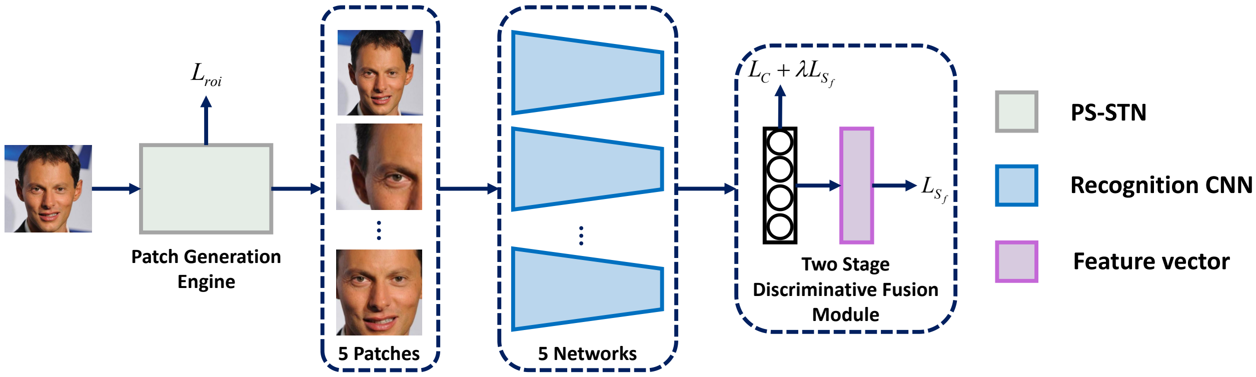

Fig. 2. The structure of our proposed end-to-end system, the STN module can adaptively search the reasonable face patches by a multi-task learning. L_{roi} , L_C and L_{S_f} represent our ROI shrunk loss, center loss and softmax loss.

and enhance the ability of single CNN model. Nevertheless, many popular deep metric learning algorithm [10], [20], [23], [29] get further performance promotion by ensembling different face patches. This proofs that patch models ensembling is an independent problem compared with metric learning problem in face recognition area. Patch models ensembling uses complementary information from different appearance of face areas and brings stability enhancement of overall performance. In this work, we concentrate on designing novel and effective framework for ensemble learning.

In most deep learning based face recognition methods, the inputs to the deep model are aligned face images during both training and testing, face alignment is performed by fitting a 2D or 3D geometric transformation between the positions of detected facial landmarks and certain predefined landmarks. [36] proposed an alignment-free system and learn the transformation matrix in an end-to-end system. Compared with these methods, our approach uses aligned face images and focus on learning complementary and discriminative face patches automatically. First, the proposed Patch Generation Engine(PGE) use Patch Search Spatial Transformer Network (PS-STN) with ROI shrunk loss to determine the location and scale of each patch. Then a Two Stage Discriminative Fusion Module (TSDFM) is proposed to make the aggregated feature from patches more discriminative and guide PGE to further search better face patches.

III. PATCH GENERATION ENGINE

In order to design an end-to-end and effective ensemble system for face recognition. Face patch selection is an important prerequisite for generating final discriminative and complementary features. In this section, we demonstrate our learnable Patches Generation Engine (PGE) which has capability of automatically searching the global optimal face patch combination.

A. Patch Search Spatial Transformer Networks

Recently, convolutional neural network (CNN) achieved impressive performance on many vision task. CNN is robust to many intra-variations such as scale, position, pose, illumination and occlusion. However, extracting the feature from a single input may not be the optimal method, because CNN might not be capable of capturing all the details from a single input image. While the complementary information from many ensembled features can take advantage of local information ignored by the network and really boost the performance [10], [20], [22], [23], [29]. Most of the existing ensemble methods in the field of object recognition like over-sample and multi-crop artificially generate different patches. While they do not take into account the relationship of different components. For unconstrained face recognition problem, the sampling mode of face patches directly affects the final performance. DeepID2 [20] proposed a simple yet effective way to tackle this situation and selected 25 best patches from 200 models greedily. This brute-force searching patch combination is flawed in that it only depend on specific evaluation dataset and may not be the optimal solution in other dataset with changing scenarios. Despite testing under same test scenarios, the fixed cropping method according the detected landmarks may not suitable for every samples ranging from large pose variations. This motivates us to find a new learnable way to alleviate these two shortcomings.

For the proposed Patches Generation Engine (PGE), we use Patch Search Spatial Transformer Networks (PS-STN) to make the patch selection process learnable. Unlike the original STN [7] which is used for searching salient object in images, our PS-STN is particularly designed for complementary facial patches selection. STN is a flexible unsupervised learning algorithm which can be easily integrated into the end-to-end system and make adaptive transform on the input images for object recognition. It has two core modules, the spatial localization network takes the input image and learns any number of image transformation parameters. These parameters can be used to implement different parametrizable transformation including translation, scaling, affine, and projective. The second module is grid sampling module which utilize original image and learned transformation parameters as input to generate transformed image.

In this paper, our PS-STN aims to simulate the region patch selection procedure under STN framework. Similar to the traditional ensemble method like DeepID2, we use aligned facial image which contains holistic face and little

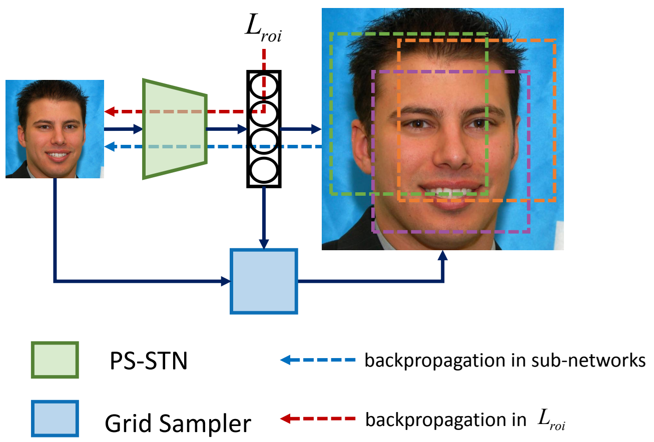

Fig. 3. The structure of our PS-STN which aims to simulate the region patch selection procedure and search the optimum solution automatically. Grid sampler is used to get cropped patches by the learned parameters. Red and blue arrows are the learning signals to guide training.

background as the original input of our framework, as shown in the left image of Fig. 2. Therefore we only need three transformation parameters $\theta_i = [s_i, t_{xi}, t_{yi}]$ to determine each face patch region to avoid deforming the original face image, where s_i is the scale of the i_{th} face region, and t_{xi} and t_{yi} are the horizontal and vertical translation parameters. A PGE generates exact n face patches, and the localization network in our PS-STN predict total $3n$ parameters. We use height and width normalized coordinates which to be in $[-1, 1]$. These three parameters can determine a square face region as follows:

$$\begin{pmatrix} x_i^{in} \\ y_i^{in} \end{pmatrix} = \begin{bmatrix} s & 0 & t_x \\ 0 & s & t_y \end{bmatrix} \begin{pmatrix} x_i^{out} \\ y_i^{out} \\ 1 \end{pmatrix} \tag{1}$$

where x_{in} and y_{in} are the coordinates mapping to the original image, x_{out} and y_{out} are the target coordinates of the regular grid in the output face patch. i represents the index of pixels in the grid, the structure of PS-STN is shown in Fig. 3.

B. Discriminative and Complementary Learning

To generate the discriminative and complementary face patches in our end-to-end system, the training strategies of PS-STN is a crucial issue. This challenges are two-fold: Firstly, capturing complementary information requires the patches to have proper distance in spatial space. Secondly, generating 'good' features requires each patch with discriminative appearance information. The learning mechanism of basic STN cannot satisfy these two requirements.

To solve these problems, we use a single branch localization network to predict the transformation parameters of all face patches simultaneously in PS-STN. This structure can consider better the relationship of all face patches than multiple localization networks. Meanwhile, we assume that patches with small overlap are more likely to contain complementary information. In order to search complementary patches and avoid predicted patches from PS-STN falling into the same region, We propose a novel ROI shrunk loss which is similar to the loss used in deep metric learning.

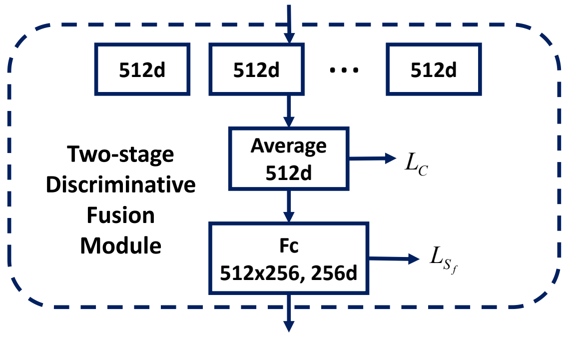

Fig. 4. The details of TSDFM

Such as triplet loss [15], contrastive loss [20] and center loss [30]. This loss further constrains the relationship of the predicted patches and make all patches adaptively transform to a more reasonable form. ROI shrunk loss is then defined as:

$$L_{roi} = \frac{1}{N^2} \sum_i \sum_j exp\left\{ -\frac{1}{2\sigma^2} \|\theta_i - \theta_j\|_2^2 \right\} \tag{2}$$

where θ_i and θ_j are the vectors composed of transformation parameters. σ is the hyperparameter that control the margin of different face patches, we set it to 0.1 in our experiment. Each of them represents a learned face patch. In this form, a face patch can be regarded as a point in the three-dimensional space. We calculate the Euclidean distance between all vectors and give all patch pairs a penalty which drops exponentially to zero as the distance between the two points increases. It is equate to vary the two face patches when they have a high overlapping. In general, the learning signal from follow CNN lead PS-STN to search more discriminative region and ROI shrunk loss can team with PS-STN to keep the complementarity of the learned patch combination, as shown in Fig. 3. These two key components constitute the proposed PGE.

IV. DEEP ENSEMBLE LEARNING FOR FACE RECOGNITION

We now introduce our end-to-end fusion module named Two Stage Discriminative Fusion Module (TSDFM), which further improve the capability of overall feature representation and guide PGE to learn better patches simultaneously.

A. Two Stage Discriminative Fusion Module

In this work, we aim to achieve end-to-end CNN architecture and generate the most discriminative fusioned feature which is superior than any single patch feature. Feature fusion is widely investigated in the hand-crafted feature era. Specifically, the feature fusion can easily be categorized as two groups: (1) feature aggregation or (2) subspace learning. Group (2) can be further classified as (i) unsupervised learning [16], [27] or (ii) supervised learning [1], [4]. However, in deep learning era, the deep feature learning is not extensively investigated. For example, the well-known DeepID2 [20] just simply concatenates all the features from different networks. This fusion is unsupervised and also not end to end, leading to less discriminative feature fusion. In the field of face recognition, the very recent work [5] proposed an end to end deep feature fusion method: neural tensor fusion. However, the dimensionality of tensor is very high, leading to difficulty of optimization.

In this work, we approach face patch feature fusion problem. In human's cognition, the most discriminative facial information of different people located in different facial region. The more crucial facial components we observe, the more confidence we have for recognizing a person. This cognitive process can be simulated in the feature space. Using single patch features can only coarsely distinguish each identity. With more patch feature aggregated, the intra-class features variations are reduced while inter-class features differences are enlarged. However, it is not easy to satisfy both two requirements of this feature fusion process. The first requirement is reducing the high-dimension of the feature vectors extracted from multi-patches and the second one is generating more discriminative feature. These need to delete redundant information and preserve critical information from extracted patches features.

To achieve this, we leverage the offline supervised fusion method in hand-crafted era and utilize the identity labels to realize a novel end-to-end supervised fusion method called Two Stage Discriminative Fusion Module (TSDFM). TSDFM breaks up the fusion process into two sepreate jobs, intra-class compacting and inter-class separating. The diffculty of each job is smaller than the whole fusion process. Therefore, we can achieve the discriminative feature generating process with low computational complexity. The structure of TSDFM are shown in Fig. 4.

In the first stage, we concentrate on increasing intra-class compacting. Specifically, we first simply average the features from all the sub-networks, then the aggregated feature is followed by a center loss [30] which simultaneously learns a center for deep features of each class and penalizes the distances between the deep features and their corresponding class centers. The loss can be formulated as:

$$L_C = \frac{1}{2}\sum_{i=1}^{m} \|x_i - c_{y_i}\| \tag{3}$$

where x_i is the fused feature in high dimensional space. c_{y_i} denotes the y_ith class center of deep features. A softmax loss is also combined with the center loss in same layer to keep training stable like [30]. After we obtain preliminary aggregated feature, in the second stage, we apply a 256D fully-connected layer to generate final aggregated feature and connected it with softmax loss which enlarges inter-class distance. This stage further reduce feature dimension and eliminate some potential noise information from the aggregated feature. Because our whole system is end-to-end, therefore, the proposed TSDFM supervises both PGE and patch sub-networks to generate the most proper feature combination. Therefore, the overall loss function of our end-to-end ensemble system is describe as:

$$L = \lambda_1 L_{Roi} + \lambda_2 L_{S_p} + \lambda_3 L_C + \lambda_4 L_{S_f} \tag{4}$$

where L_{Roi} is our ROI shrunk loss, L_{S_p} is softmax loss used for each patch sub-network, L_C and L_{S_f} is the center loss and softmax loss used at our TSDFM. λ is the weight of loss function and is set to 1 in all of our experiments.

V. EXPERIMENTS

In this section, we first introduce our implementation details. Then We evaluate the performance of our system by doing some ablation experiments and comparing with the state-of-the-art methods.

A. Implementation Details

Data Preprocessing: First, we use MTCNN [35] to align the face images and cropping the region which contains the whole face as shown in Fig.2. Then we resize every image to 120 $\times$ 120. The images are horizontally flipped for data augmentation during training. We normalize each pixel of training images to -1.0-1.0. To be fair, we do not mirror the test image in all of our experiments.

Network architectures: In PS-STN, we use a CNN with 4 convolutional layers and two FC layers as the localization network. Each recognition CNN in Fig.2 is a 10 layer ResNet [3] to circumvents the problem of performance saturation.

Training Strategy: Our system is implemented based on PyTorch [14] with a batch size of 128. The base learning rate in our experiments is 0.1 except PS-STN which multiplies the base learning rate by 10^{-3} for stability. The system was trained for 55k iterations with SGD and we reduce the learning rate by a factor of 10 after 20k, 40k, 50k.

Datasets: In our experiments, we only use CASIA-Webface [34] datasets for training and test performance on the LFW [6] dataset and YTF [31] dataset. The CASIA-WebFace dataset is a large-scale dataset containing about 10,575 subjects and 500,000 images from Internet. This unconstrained dataset has accelerated the development of face recognition in the wild.

LFW dataset contains 13,233 web-collected images from 5749 different identities, with large variations in pose, expression and illuminations. we follow the standard evaluation protocol and test on 6,000 face pairs.

YTF dataset consists of 3,425 videos of 1,595 different people, with an average of 2.15 videos per person. The clip durations vary from 48 frames to 6,070 frames, with an average length of 181.3 frames. Again, we follow the unrestricted with labeled outside data protocol and report the results on 5,000 video pairs in Table 2.

B. Ablation Study

In this section, we run a number of ablations to analyze the effectiveness of our proposed system. we evaluated different patches selection strategies and different fusion methods to demonstrate the effectiveness of our end-to-end ensemble framework with the novel PGE and TSDFM.

Firstly, we investigate the performance of different patches selection strategies. As shown in Table I, we first train a CNN with the original aligned face images as our baseline. These aligned face images also serve as the inputs of our ensemble system to generate different face patches. Then we

TABLE I
COMPARISON OF DIFFERENT PATCH SELECT METHODS

Num	Selection	Fusion	LFW mAC(%)
1	-	Concat	97.40%
5	Random	Concat	97.52%
5	Greedy	Concat	97.78%
5	PS-STN	Concat	98.03%
5	PGE	Concat	**98.35%**

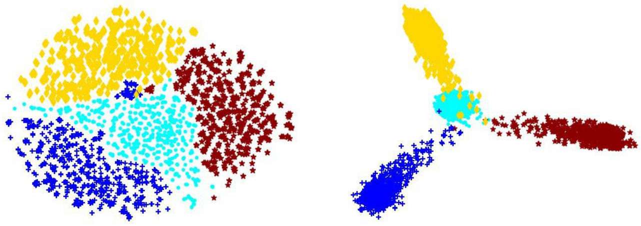

Fig. 5. Visualization of the impact of ROI shrunk loss. The left one shows the result without ROI shrunk loss, the right one are with ROI shrunk loss. Each color represents the feature extracted from one patch sub-network.

respectively apply artificial selection method and learnable selection method with the same concatenation fusion method. We trained 5 sub-networks in every experiment. It should be noticed that the original aligned face images are used in all experiments. Therefore, the patch selection methods only need to crop four patches from the original face image. As for the artificial selection method, we can easily observed that greedy searching method is outperform than the random searching strategy, but it is still not better than the learnable patch searching method. Our PGE brings verification rate by 0.57%. The reason is that the learnable searching methods can adaptively adjust the face patch in a small range during the testing phase and the fixed cropping method based on detected landmarks may not suitable for every samples ranging from large pose variations. Adding ROI shrunk loss to our PS-STN brings 0.32% gains, that indicates the importance of the complementary information.

In order to demonstrate the effectiveness of our Two Stage Discriminative Fusion Module (TSDFM), we investigate the performance of different feature fusion methods. From Table II, we can see that the unsupervised fusion methods like average fusion (AVG), concatenation (Concat) perform worse than other supervised fusion methods. Meanwhile, we keep the architecture of TSDFM but use single center loss (Intra) or Softmax loss (Inter) for supervised fusion. The results show that our TSDFM which considers both intra-class distance and inter-class distance in different layers outperforms Intra and Inter (99.03% vs 98.82% and 99.03% vs 98.65%). We also conduct an experiment that only uses TSDFM for training and uses the simple concatenation fusion method during testing phase (TSDFM-). We can see TSDFM- outperforms Concat (98.80% vs 98.35%), which demonstrates that TSDFM also has the ability of leading PGE to learn better face patches.

TABLE II
COMPARISON OF DIFFERENT FUSION METHOD

Selection	Fusion	LFW mAC(%)
PGE	Avg	97.82%
PGE	Concat	98.35%
Greedy	TSDFM	98.32%
PGE	Inter	98.65%
PGE	Intra	98.82%
PGE	TSDFM-	98.80%
PGE	TSDFM	**99.03%**

TABLE III
FACE VERIFICATION ACCURACY ON LFW AND YTF. WE DIRECTLY COMPARE TO RESULTS TRAINED ON CASIA-WEBFACE, NOTE THAT THE METHODS SHOWN IN THE UPPER PART USE MORE OR PRIVATE DATA.

Model	Images	Acc.(LFW)	Acc.(YTF)	Layers.	Nets
FaceNet [15]	200M	**99.63**	95.1	14	1
DeepFace [26]	4.4M	97.35	91.4	7	1
MultiBatch [25]	2.6M	98.20	-	12	1
VGG [13]	2.6M	99.13	97.3	16	1
DeepID2 [20]	203K	99.15	93.2	5	25
DeepID3 [21]	300K	99.53	-	22	25
CASIA [34]	494k	97.30	-	11	1
MFM [32]	494k	98.13	-	29	1
N-pairs [19]	494k	98.33	-	11	1
CenterFace [30]	494k	99.00	94.9	29	1
SphereFace [10]	494k	**99.42**	-	64	1
Greedy+Concat	494k	97.78	91.8	10	5
Ours(PGE+TSDFM)	494k	**99.03**	**94.5**	10	5

C. Effectiveness of ROI shrunk loss

For better understanding the complementary face patches learned by our system and evaluate the effectiveness of the proposed ROI shrunk loss. The direct purpose of using this loss is to make the regions of the face patches different. We also plot the high dimensional feature of each patch on 2-D surface for visualization by using t-SNE proposed in [12]. From Fig. 5, we can observe that the deeply learned features of single person's different patches are separable. Using proposed ROI shrunk loss joint with other loss not only reduce the overlap of different patches in spatial space but also make each patch features more discriminative. Without this loss, the learned face patches would have a big overlap and bring considerable redundant information during the fusion phase.

D. Comparison with Existing Methods

Table III presents results for face verification. Our system achieves competitive accuracy among the other models trained on CASIA-WebFace. MFM [32] trains with softmax classification loss. CASIA [34] trains with a combination of softmax loss and contrastive loss. CenterFace [30] and SphereFace [10] are two popular metric learning methods. CenterFace uses a more deep CNN with 29 layers and SphereFace uses 64 layer CNN, while the depth of the networks in our system are all 10. In this paper, we only concentrate on proposing the learnable ensemble face recognition system which is much superior than the existing face recognition ensemble methods. Therefore, we only use a light-weighted 10 layer CNN trained with softmax loss.

With the deployment of more powerful CNN architecture and loss function, we believe our system has a lot of potentiality to be improved in the future. We also evalute our method on YouTube Faces (YTF) benchmark which is more difficult than LFW due to the low-quality images. Our PGE+TSDFM significantly outperforms the baseline (Greedy+Concat), 94.5% vs 91.82%, as shown in Table III.

VI. Conclusions and Future Works

We proposed an end-to-end face recognition framework in which the complementary and discriminative face patches can be learned automatically. Relying on the fact that image crops can actually be obtained from an aligned face image via translation, scale and cropping operations, we proposed a novel Patch Search Spatial Transformer Network (PS-STN) and ROI shrunk loss in our architecture to perform the face patches selection process. In order to extract more discriminative and complementary features from the learned face patches. we adapt a ROI shrunk loss which enlarges the distance of learned features in spatial space and feature space. These two modules constitute our Patch Generation Engine (PGE). We also propose a Two stage discriminative fusion module (TSDFM) to aggregated the feature. In our future work, we will focus on the ensemble method of face recognition problem continually and design new algorithm to enhance substantial performance.

VII. Acknowledgments

This work is supported by the Natural Science Foundation of China (Grant No. 61772527, 61806200).

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. Technical report, Yale University New Haven United States, 1997.

[2] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):328–340, 2005.

[5] G. Hu, Y. Hua, Y. Yuan, Z. Zhang, Z. Lu, S. S. Mukherjee, T. M. Hospedales, N. M. Robertson, and Y. Yang. Attributeenhanced face recognition with neural tensor fusion networks. In *Proc. Int. Conf. Comput. Vis.(ICCV)*, 2017.

[6] 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. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[7] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.

[8] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li. Learning multi-scale block local binary patterns for face recognition. In *International Conference on Biometrics*, pages 828–837. Springer, 2007.

[9] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.

[10] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 1, 2017.

[11] W. Liu, Y. Wen, Z. Yu, and M. Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, pages 507–516, 2016.

[12] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[13] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[14] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[15] 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*, pages 815–823, 2015.

[16] J. Shawe-Taylor, N. Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[17] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[19] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1857–1865, 2016.

[20] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014.

[21] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[22] 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*, pages 1891–1898, 2014.

[23] 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*, pages 2892–2900, 2015.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[25] O. Tadmor, Y. Wexler, T. Rosenwein, S. Shalev-Shwartz, and A. Shashua. Learning a metric embedding for face recognition using the multibatch method. *arXiv preprint arXiv:1605.07270*, 2016.

[26] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[27] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000.

[28] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[29] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 1, 2018.

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

[31] L. Wolf, T. Hassner, and I. Maoz. *Face recognition in unconstrained videos with matched background similarity*. IEEE, 2011.

[32] X. Wu, R. He, and Z. Sun. A lightened cnn for deep face representation. In *2015 IEEE Conference on IEEE Computer Vision and Pattern Recognition (CVPR)*, volume 4, page 5, 2015.

[33] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[34] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[35] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[36] Y. Zhong, J. Chen, and B. Huang. Toward end-to-end face recognition through alignment learning. *IEEE signal processing letters*, 24(8):1213–1217, 2017.