IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 20, 2020, accepted December 31, 2020, date of publication January 6, 2021, date of current version January 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049639

Attention-Modulated Triplet Network for Face Sketch Recognition

LIANG FAN, XIANFANG SUN, AND PAUL L. ROSIN
School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K.
Corresponding author: Liang Fan (fanl7@cardiff.ac.uk)

ABSTRACT In this paper, a novel triplet network is proposed for face sketch recognition. A spatial pyramid pooling layer is introduced into the network to deal with different sizes of images, and an attention model on the image space is proposed to extract features from the same location in the photo and sketch. Our attention mechanism builds and improves recognition accuracy by searching similar regions of the images, which include abundant information in order to distinguish different persons in photos and sketches. So that the cross-modality differences between photo and sketch images are reduced when they are mapped into a common feature space. Our proposed solution is tested on composite face photo-sketch datasets, including UoM-SGFS and e-PRIP dataset, and achieves better performance than the state-of-the-art result. Especially for Set B in UoM-SGFS dataset, the accuracy is higher than 81%.

INDEX TERMS Attention module, triplet network, attention mechanism, cross-modality recognition, shared-weight network.

I. INTRODUCTION

Photo-sketch image recognition is a type of cross-modality recognition whose aim is to take a given face sketch and search for a corresponding photo from a face photo dataset. Face sketch images are often the only description of suspicious persons when their appearance was not captured on videos. This technology also supports to look for missing people.

Because photos and sketches have different appearances, it is not possible to directly use face photo recognition methods for face sketch recognition. Tang and Wang [1] proposed to reduce the modality discrepancy by synthesising a photo from the sketch. After that, feature descriptor [2] and mapping methods [3] are used to reduce the modality gap. These methods have a high recognition accuracy for some hand-drawn face photo-sketch datasets, and some can even achieve 100% recognition [4]. However, in practice, people more typically use composite face sketches rather than hand-drawn face sketch images, and in these cases, the above-mentioned methods do not obtain satisfactory performance since photos have more substantial geometric and photometric differences compared to composite face sketches than hand-drawn sketches.

Recently, deep learning techniques were introduced into face photo-sketch recognition to obtain similar features in a common space. Some of the methods [5], [6] transfer the images into the same modality by generating synthesized pseudo images from sketches using GAN networks. [7]–[9] try to extract the deep image features from multi-channel neural networks, such as Siamese networks and triplet networks. In spite of the abundant features extracted using deep learning methods, the limited number of datasets and the weak convergence of loss functions results in unsatisfactory recognition. In Siamese networks, the extracted features of photos and their corresponding sketches are inconsistent because of the deformation of sketches with respect to the photos.

In this paper, we propose an attention module and a spatial pyramid pooling layer in a triplet network to improve the efficiency and accuracy of face photo-sketch recognition. The attention module is used to learn the related features at similar locations from cross-modality images. It consists of two attention blocks: One attention block acts on both the face photo and the sketch to generate the feature maps, and another attention blocks act on the photo and the sketch to focus on the location of the facial features. Both blocks share the same structure, and it related to the photos is trained first because large training datasets are available. Then, the attention block for sketches is trained using fine tuning and a smaller sketch training dataset to adapt the photo attention block.

To improve the similarity between a face photo and its corresponding face sketch, the photo feature map from the attention module is adopted to the sketch image to extract the same attribute information. In order to avoid the influence of the noise and distortion after cropping the images, instead of a full connected layer in the convolution neural network,

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

a spatial pyramid pooling layer is used to deal with the input images of random sizes without pre-processing methods, such as cropping and scaling. To optimize the proposed triplet network, the triplet loss function is used to compare the feature distance between the input images using the chi-square distance. The experiments show that the performance of our proposed method achieves better results than the state-of-the art result on composite face photo-sketch recognition.

The contributions of this paper are followed as:

1, We develop a triplet network combined with an attention module and a spatial pyramid pooling layer. The parameters of each channel are shared to generate the same encoding rules for extracting features before the attention module.

2, We design an attention module which consists of two attention blocks to focuses on extracting similar shape features from different modalities of images (photo and sketch).

3, The pooling layer is introduced to reduce the effect of image noise and deal with different sizes of input images.

II. RALATED WORK

A. TRIPLET NETWORK

The aim of the triplet network is to distinguish different classes of images and identify the same class of images after comparing the feature distances from each CNN channel. In the earliest model, Hoffer and Ailon [10] proposed to compare the feature distances after mapping the images into an embedding space. This network consists of three feed-forward networks to reduce the sensitivity to calibration in the sense that the notion of similarity vs dissimilarity requires context. A distance margin is used in the loss function to set a clear boundary between the inter-class samples and the intra-class samples. In FaceNet [11], a triplet loss function is proposed to minimize the feature distance between the face photos of the same person and maximize that of different persons. The distance between the anchor samples and the positive samples is closer than that between the anchor samples and the negative samples after optimization. The advantage is that the loss function utilizes the difference of the samples to distinguish their details. To reduce training time, in deep face recognition [12], the strong supervised character of the SoftMax function is utilized to increase the fitting speed of the model. After training the model, the features are extracted from the trained model and the distance between the sample features is optimised using the triplet loss function. Hermans *et al.* [13] introduced a constraint on the triplet loss function to keep the distance between the samples of the same identity closer than those of different identifies. This loss function focuses on increasing the recognition accuracy for complex images and the generalization ability of the model for fine-grain image recognition and cross-modality image recognition.

B. ATTENTION MECHANISM

The attention mechanism selects a state similar to a given one from some set of states in the network, and then performs subsequent information extraction [14], [15]. The aim is to compare the similarity between the vector collection $v_1, v_2, \ldots v_n$ and assign big/small weight values to those with high/low similarity, respectively.

Mnih *et al.* [16] proposed an attention mechanism based on a reinforcement learning model. It extracts the information from a picture or a video, and selects a range of regions or locations, which then processed at high resolution. The recurrent model is allowed to train a given task directly using the past information and mission requirement. It not only extracts the features from the whole image, but also extracts the necessary feature using the relationship between the image pixels. Vaswani *et al.* [17] utilized a self-attention module to capture the related global information using the hidden state from the source input and the target input data. Compared with the traditional attention mechanism, the relationship between the source and the target data is used to obtain the dependency of the source and the target data. Yin *et al.* [18] applied an attention mechanism on a convolution neural network. The CNN model selects the downsampling layer to keep the scale and spatial invariance of the input images. However, the pre-selected fixed size is limited to adapt to the deformation, and so the feature map of the images cannot reflect the deformation of an image and the whole image's feature. In the spatial transformer network [19], the localisation network uses a subnetwork that is a component of the CNN model to generate the spatial transformation parameters, which can transfer the input map to the expected output map. The learned spatial transformation network will automatically extract the local data features from the attention area, and eliminate deformation on the target image by applying a reverse spatial transformation of the image. SENet [20] built the correlation between feature channels to intensify the important feature for recognition. The core idea of SENet is to learn feature weights based on the loss function through the network and get the importance of each feature channel automatically. The useful features used in current tasks are increased and better results are achieved according to this importance of features. The network extracts the spatial information as a 'global descriptor', then two full connected layers generated a feature map. Finally, the feature map multiplied with the original space after the global average pooling to recalibrate the output feature map. However, it does not reflect the meaning of attention in the spatial dimension. The similar features are related to each other without distance. Woo *et al.* [21] applied the attention mechanism on the channel and the spatial dimensions separately to improve the feature extraction ability of network models without increasing the amount of calculation and parameters significantly. The channel and the spatial attention modules work on the input feature maps sequentially to generate refined features.

III. MODEL DESIGN

Our face photo-sketch recognition system consists of three parts. The first part is a triplet network for few-shot recognition. The second is an attention network which is introduced

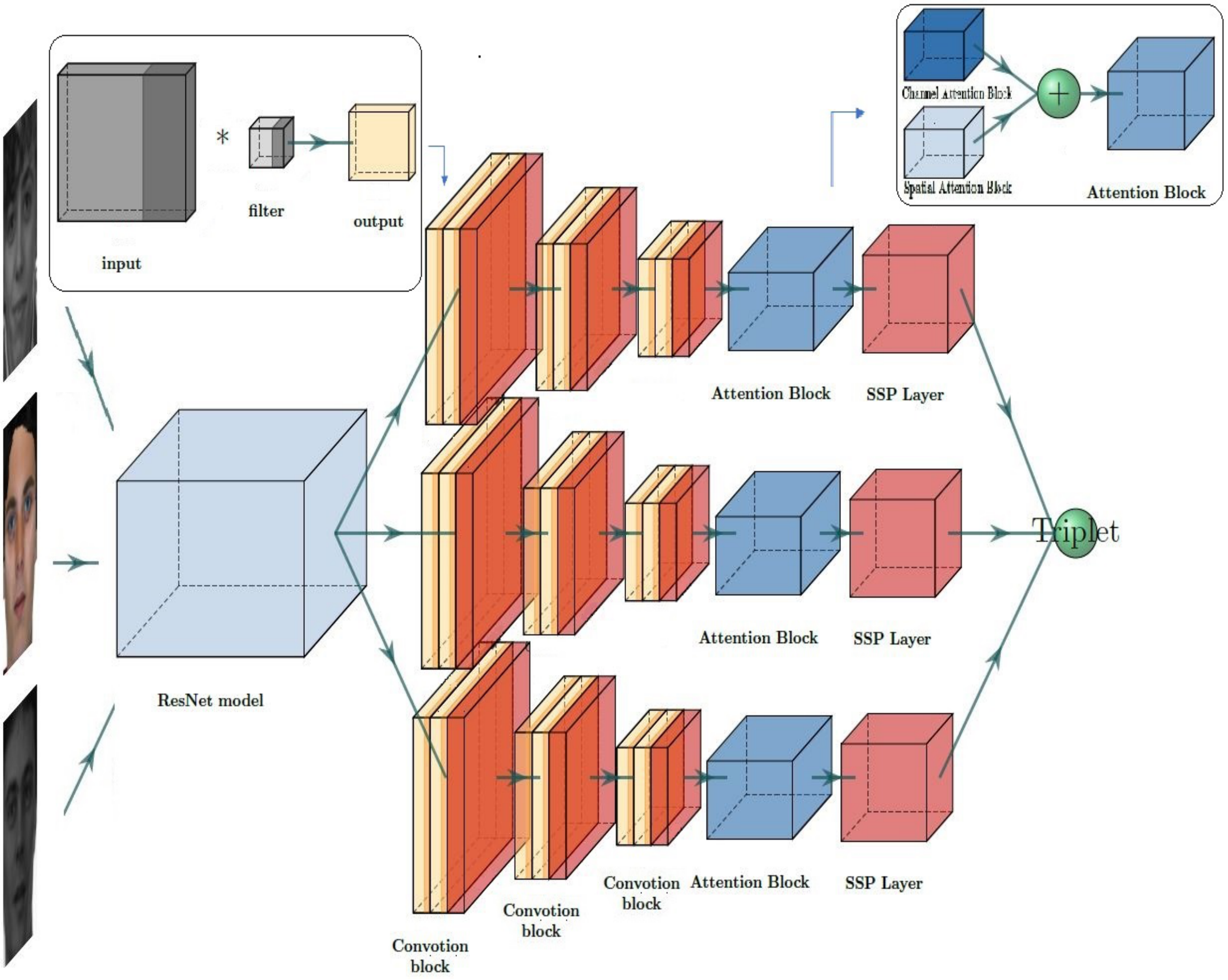

FIGURE 1. The structure of the triplet network and details of each channel. The features are extracted from the fourth convolution block from the ResNet model. The structure of each convolution block consists of two convolution layers (yellow) and a max-pooling layer (red). The kernel size of each convolution layer is 3 ∗ 3 and the stride size is 2. The input image size is $n * n * 3$, and the output size is $\frac{n+1}{3} * \frac{n}{3} * 2$.

to extract similar feature vectors from both the photo and the sketch images. The third is a pooling (SPP) layer included to avoid information loss due to the fixed size of the input images. The proposed triplet network consists of three branches of neural networks, as is shown in Figure.1. Each channel is constructed by three convolution layers, an attention block, and an SPP layer. In this triplet network, two of the input images are the sketch and the photo images of the same person, and the third input is the face sketch of a different person. The feature of face photos and face sketches have full performed for recognition, the channel attention model gives the different weights for the features which extracted from the previous convolution layer. Then, the different generated mechanisms lead to that the extracted features are 'asymmetry'. However, the ratio of the distance between each attribute (eyes, nose, mouth, and so on) on face images keeps invariable. The contextual information from our spatial attention block focus on the unique facial characters. We utilize a shared-weights triplet network to extract a similar feature from the photo and the corresponded sketch, excluding spatial attention block. The unshared spatial attention block utilizes the contextual feature to find the relationship between the photo and the sketch. The image feature maps will be extracted from each channel and used in a triplet loss function to minimize the feature difference between the image pairs of a same person and maximize that of different persons. The triplet loss function is defined as

$$L = \sum\nolimits_i^N \left[\left\| f\left(x_i^a\right) - f\left(x_i^p\right) \right\|_2^2 - \left\| f\left(x_i^a\right) - f\left(x_i^n\right) \right\|_2^2 \right] \tag{1}$$

where N is the number of input image triplets, $f(\cdot)$ is the feature map function, x_i^a is the anchor face photo image, x_i^p is a positive image, i.e., the face sketch of the same person as in the anchor image, and x_i^n is a negative image, i.e., the face photo of a different person. We use the chi-squared distance to measure the feature differences of different images. Compared with Euclidean distance, chi-square distance utilizes the similarity between the same location's features to optimize our model.

A. THE ATTENTION NETWORK

Because the features extracted from photos and sketches using the trained shared-weight network are different for each attribute of the face. Such as, the sketch which draws by lines choose a larger size filter to extract the features that represent the characteristics of the face sketch. The large size filter is difficult to extract similar features that focus on the characteristics of face sketch from the corresponded photo. We design an attention module to extract features from effective regions of the images to increase the recognition accuracy. The attention mechanism is used in each channel of the triplet network. There are two parts in the attention module. One is designed to obtain the relationship between the images of each channel, and focuses on extracting the shape of the input images. And the other focuses on extracting the features of similar locations.

The proposed attention module consists of two blocks. The convolution operations lead to a local receptive field. The features corresponding to the pixels with the same location of face photo and face sketch may have some differences. The difference introduces inconsistencies between the intra-class images and the inter-class images. Our method can adaptively focus on the features of the same position between face photo and face sketch using the attention module to enhance the feature representation for recognition. The attention module in our network pays more attention in the important parts of the images, which is useful for matching between the face photo and face sketch images. It includes a channel attention block and a spatial attention block to extract the edge feature and texture feature from the input images. Finally, the feature map from the spatial pyramid pooling layer is fed into a fully connected layer with L2 normalization. The input of the attention module is the feature map F_{conv3}, which is extracted from the third convolution layer. The attention feature map which contains the channel attention feature map $F_{channel}$ and spatial attention feature map $F_{spatial}$ is as follows:

$$F_{attention} = [F_{channel}, F_{spatial}] \tag{2}$$

B. THE DETIAL OF ATTENTION BLOCKS

Human utilizes the facial attributes' shape to recognize people from a set of face images dataset. Therefore, artists draw the face sketch based on the facial attributes' shape. Due to the limited geometries that can be used to describe the human face shape and facial characters on a sketch, the fine-grained features of each facial attribute' shape and relationship between each facial attribute' features are the key clue for recognition. Thus, two convolution neural layers extract the fine-grained features. Meanwhile, we use two types of pooling layer to obtain the edge information of features. One is a global pooling layer reduces the extracted features' dimensions. Another is a max-pooling layer calculates the features' location.

As illustrated in Figure.2, the attention block of photo images utilizes the intra-channel relationship between the extracted features from the convolution layer to represent meaningful features for recognition. The channel attention map is computed as:

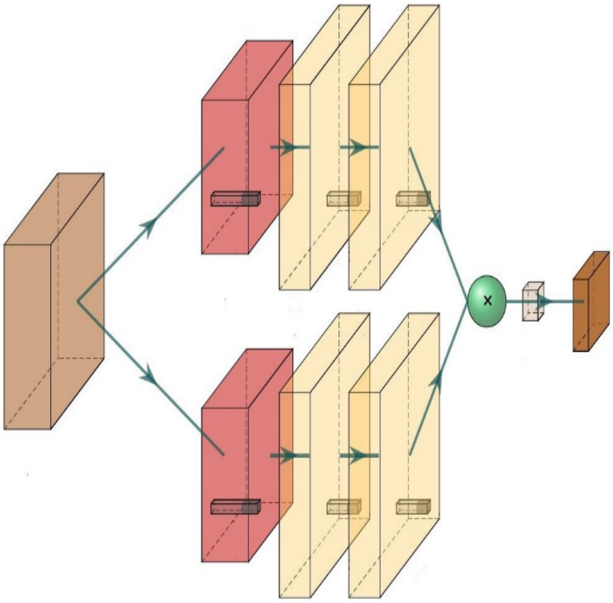

FIGURE 2. The structure of the channel attention block which extracts the intra-channel features from photos and sketches. This attention block consists of two channels. Except for the last layer of each channel, two convolution layers (yellow) and one max-pooling layer (red) are included in the module. The kernel size is 1 ∗ 1 for each convolution layer and pooling layer.

$$F_{edge} = GloPool(conv(conv(F_{conv3}))) \tag{3}$$
$$F_{texture} = MaxPool(conv(conv(F_{conv3}))) \tag{4}$$
$$F_{channel} = Sigmoid(F_{edge} \times F_{texture}) \tag{5}$$

To compute the edge features F_{edge}, global average pooling is used to keep the edge information on the feature map. We fed F_{conv3} into two convolution layers followed by a global average pooling layer to get a feature vector F_{edge}. Meanwhile, a max-pooling layer is used to extract the texture feature $F_{texture}$ with F_{conv3} passing through two same convolution layers. Then matrix element-wise multiplication between F_{edge} and $F_{texture}$ is performed to form an integrated channel feature $F_{channel}$ that contains the edge and texture features. Finally, an *ELU* activation function is used to obtain a channel feature without image noise.

Different from the first attention module which is used to obtain the relationship between different cross-modalities' images, an attention module pays more attention to the position of a picture. We use two types of modules to act on the photo image and the sketch image to extract similar location's feature, as shown in Figure.3. The spatial modules adopt the stochastic pooling layer [22] combined with the overlapping pooling strategy [23]. The pooling map of the stochastic pooling layer has the probabilities p_i for each pooling region j that is computed by a multinomial distribution after nominalizing the feature map a_i as:

$$p_i = \frac{a_i}{\sum k \in R_j a_k} \tag{6}$$

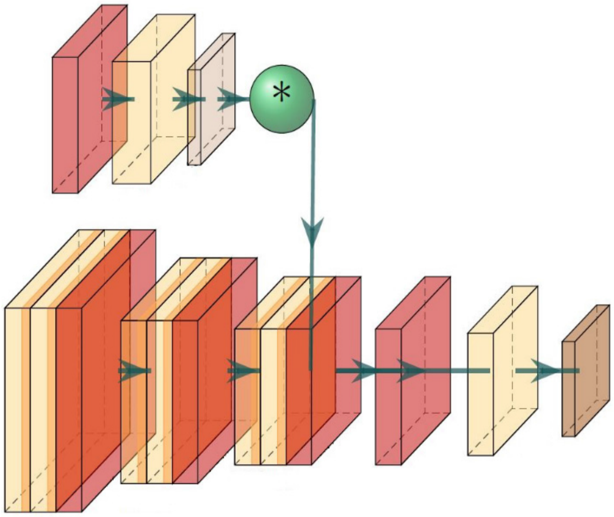

FIGURE 3. The structure of the spatial attention module for the photo and sketch images. The attention module consists of a stochastic pooling layer and a convolution layer. The kernel size of the stochastic pooling layer is 1∗1, and the kernel size of the convolution layer is 3 ∗ 3 with padding size 1. Following the convolution layers, the sigmoid function is used as an activation function.

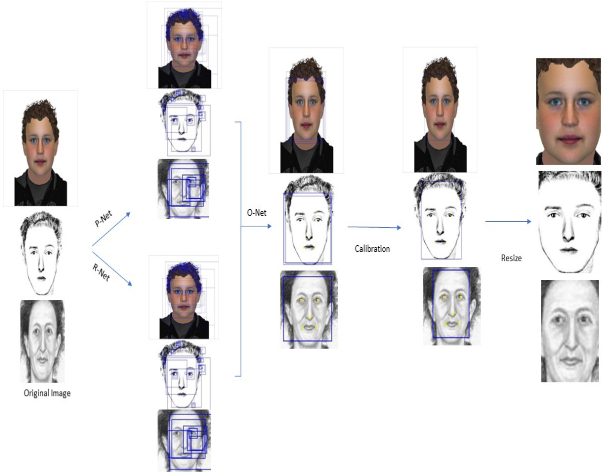

FIGURE 4. The pre-process results using MTCNN model on face sketch images, including the colour component sketch, line sketch and hand-drawn sketch.

The sample will be selected randomly with the multinomial distribution p_i from each pooling window. Different from classic pooling strategy, overlapping regions exists between the adjacent pooling windows in this overlapping pooling strategy to avoid loss of information from the input feature map. This method fuses the multi-level features using sparse sampling to increase the robustness of the target deformation. According to the principle of this strategy, the stride s of each filter window must be larger than the pooling window size z in the stochastic pooling layer. We obtain a pooling map F_{pool_photo} from the feature map F_{conv3_pool} which is extracted from the third convolution layer. One advantage of this method is that the edge information and texture information are extracted without distortion. Another advantage is that the image noise is reduced. Then F_{pool_photo} is fed into a convolution layer to generate an attention map $F_{spatial\_photo}$ which is computed as:

$$F_{spatial\_photo} = Conv(F_{pool\_photo}) \tag{7}$$

$$F_{pool\_photo} = [AvgPool\left(F_{conv3\_pool}\right), StochPool\left(F_{conv3\_photo}\right)] \tag{8}$$

where [•, •] means concatenation of two feature vectors.

The triplet network is designed to use shared weights to extract similar features from each channel of the network. However, the convolution layer ignores the inter-relationship from each patch of the sketch. We utilize the edge feature vector of the photo, which is generated from photo spatial attention module, to focus on the same position and extract more similar features from the sketch's channel attention $F_{channel\_sketch}$. The structure of the spatial attention module for sketch is as follows:

$$F_{point\_sketch} = F_{spatial\_photo} \times F_{conv3\_sketch} \tag{9}$$

where × denotes the elementwise multiplication between the photo spatial attention and the sketch feature map, so as to extract the features from similar positions in both images. The sketch edge feature vector $F_{spatial\_sketch}$ is extracted using the same method as the photo spatial attention layer to allocate more weight to the image. After that, a spatial feature map for a sketch image is obtained after calculating the correlation using a sigmoid layer.

$$F_{spatial\_sketch} = \sigma(Conv\left(F_{pool\_sketch}\right)) \tag{10}$$

$$F_{pool\_sketch} = [AvgPool\left(F_{point\_sketch}\right), StochPool\left(F_{point\_sketch}\right)] \tag{11}$$

C. THE DETAIL OF SPATIAL POOLING LAYER

We adopt a spatial pyramid pooling layer [24] after the attention module, instead of the fully connected layer in the original triplet network. Spatial pyramid pooling allows not only inputs of arbitrary aspect ratios, but also arbitrary scales. One purpose of using spatial pyramid pooling is to avoid information loss from cropping. The second is that different ratio features can be extracted from the feature map of the attention module to increase the robustness of our method. The input feature map after the attention module retains the original features from the last convolution layer. Re-extracted features from the sketch with fewer features through down sampling layer. Then using the small features extracted from the photo by deconvolution layer to fuse the three types' features. This fusion method utilizes abundant features to eliminates the influence of the inconsistency of effective feature information due to the large difference between sketch and photo. Otherwise, average pooling is applied on each sub-window. After average pooling for each sub-window, the dimensions of the feature maps for each sub-windows are the same. The output feature map is composed of the feature maps from all sub-windows.

IV. EXPERIMENT

To evaluate our model, we build a triplet network using the PyTorch framework and test it on the component face photo-sketch dataset (UoM-SGFS and e-PRIP dataset and hand-drawn face photo-sketch dataset (CUFSF dataset).

A. PRE-PROCESS METHOD

The key points of each image have deviation from each other, and the cropped face images do not keep the same size. To keep the whole information of face image and reduce image noise, MTCNN [25] is utilized to detect the location of each face image and rotate the image. In our experiments, the performance of MTCNN was found to be satisfactory for both photos and sketches. After pre-processing, the key points of the face image and its sketch are aligned.

B. IMPLEMENTATION DETAILS

The input for our model consists of three images, each of which will enter into the corresponding channel. For the first and the last channel, the input images are face photos, while the input of the middle channel is the face sketch of the same subject that is input to the first channel. For training all datasets, we build a model using ResNet as a pre-trained model, together with the attention module and SPP layer as described above. The initial learning rate is set as 6e-5 with the Adam optimizer.

C. EVALUATION ON UoM-SGFS DATASET

UoM-SGFS dataset [7] is a large face photo-sketch dataset, containing 300 color face photos from the Color-FERET dataset and 300 color face sketches generated using EFIT-V software. The face attributes of the sketches in set B are modified to increase the similarity between the face photo and face sketch than the sketches in set A. We test the model on both the UoM-SGFSA and UoM-SGFSB datasets.

We compare the proposed method with different approaches, as well as several state-of-the-art methods on the UoM-SGFSA dataset for rank-1 accuracy. From the feature map generated by class activation mapping [26], we can see that the weights of our network focus on the parts of the face photos and their corresponding sketches. After using the channel attention module and spatial attention module, more similar features are extracted from the same position of the face photo and the sketch. As is shown in Table 1, the FaceNet model obtains 45.50% using the shared parameter network. The accuracy increases to 66.70% for our model which combines the attention module and SPP layer with Resnet-34 as the pre-trained model.

Compared with Set A in the UoM-SGFS dataset, the face photo and corresponding sketch are more similar in Set B. The metrics of the extracted features from photos and their correspond sketches in Set B are closer using the attention module than Set A. Table 2 shows that the accuracy for Set B using our model exceeds 81%, while the accuracies of the others are less than 75%.

D. EVALUATION ON E-PRIP DATASET

We also test the attention module on another component face sketch dataset [28] whose sketches are generated based on the AR dataset using FACE software. Table 3 shows the experimental results. It can be seen that although the top-1 accuracy of our method is reduced to 58.85%, it is still the highest accuracy among the state-of-the-art methods.

TABLE 1. Experimental results on UoM-SGFSA dataset.

Method	Top-1 accuracy	Top-10 accuracy
FaceNet	45.50%	50.70%
Triplet net +SPP Layer	47.30%	55.20%
Triplet net + channel Attention block+ SPP Layer	53%	70.58%
Triplet net + Spatial Attention block +SPP Layer	52.47%	73.19%
Triplet net + Attention blocks +SPP Layer	66.75%	90.46%
[27]	64.80%	92.13%
[28]	31.60%	66.13%

TABLE 2. Experimental results on UoM-SGFSB dataset.

Method	Top-1 accuracy	Top-10 accuracy
FaceNet	52.00%	80.10%
Triplet net +SPP Layer	53.11%	82.20%
Triplet net + channel Attention block+ SPP Layer	74.91%	86.47%
Triplet net + Spatial Attention block +SPP Layer	70.16%	84.90%
Triplet net + Attention blocks +SPP Layer	81.25%	91.56%
[27]	72.53%	94.80%
[28]	52.17%	82.67%

E. EVALUATION ON HAND-DRAWN FACE PHOTO-SKETCH DATASET

The component face sketches generated by the software are close to real-life forensic sketches. However, because the options for face attributes in the software are limited, the recognition accuracy is not high, even though the attention module and the SPP layer are added to weight more on important and useful parts of the images. We also tested our model on the hand-drawn sketches [29]. Those sketches are drawn by artists based on real photos, and the similarity of the sketches to the photos is much closer than that of the component sketches. However, the dataset contains only 188 image pairs, so the amount is too small for training. We have used data augmentation to increase the number of

TABLE 3. Experimental results on E-PRIP dataset.

Method	Top-1 accuracy	Top-10 accuracy
FaceNet	50.20%	56.70%
Triplet net +SPP Layer	51.00%	60.10%
Triplet net + + channel Attention block+ SPP Layer	55.00%	75.55%
Triplet net + Spatial Attention block +SPP Layer	54.30%	72.14%
Triplet net + Attention blocks +SPP Layer	58.85%	84.60%
[27]		82.80%
[28]	54.90%	80.80%
[29]	52%	60.20%

TABLE 4. Experimental results on CUFS dataset.

Method	Top-10 accuracy
FaceNet	75.10%
Triplet net +SPP Layer	76.71%
Triplet net + channel Attention block+ SPP Layer	79.15%
Triplet net + Spatial Attention block +SPP Layer	77.36%
Triplet net + Attention blocks +SPP Layer	89.60%
[31]	92.56%

the training data. For testing, the top-1 accuracy of our model is higher than the FaceNet model and obtains 84.27%. Table 4 shows the top-10 accuracy comparison of our methods with some others. It can be seen that the result of our model is more than 8.8% better than the state-of-the-art method [30]. Our method achieved better results because the triplet network with attention can extract similar shape features while mitigate their difference caused by the difference of image modalities.

V. CONCLUSION

We presented a novel approach to enhance the recognition accuracy for face photo-sketch datasets. We built a triplet network architecture with a triplet loss function layer to learn the feature representation. To extract features that are correlated between face photos and face sketches, we introduced an attention model to focus on the same position features using a channel attention block and a spatial attention block. In addition, an SPP layer was introduced to extract the features from image blocks of different scales, and to reduce the influence of distortion and noise from the input images. To verify the model's effectiveness, we tested two kinds of face photo-sketch datasets for recognition. For component face photo-sketch datasets, the performance of our model is the best among the popular state-of-the-art methods. We improve the recognition accuracy from 64.80% [27] to 66.75% using our method. Meanwhile, the recognition rate of our method achieves 58.85% on e-PRIP dataset. The highest accuracy we obtained from the three datasets is 81.25%. We utilized the hand-drawn face photo-sketch dataset CUFS to verify that the corresponding features between face photos and hand-draw sketches are closer than between face photos and component face sketches.

REFERENCES

[1] X. Tang and X. Wang, "Face sketch recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 50–57, 2004.
[2] B.-S. Oh, K. Oh, A. B. J. Teoh, Z. Lin, and K.-A. Toh, "A Gabor-based network for heterogeneous face recognition," *Neurocomputing*, vol. 261, pp. 253–265, 2017.
[3] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 593–600.
[4] H. K. Galoogahi and T. Sim, "Inter-modality face sketch recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2012, pp. 224–229.
[5] Y. Güçlütürk, U. Güçlü, R. van Lier, and M. A. van Gerven, "Convolutional sketch inversion," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 810–824.
[6] Q. Guo, C. Zhu, Z. Xia, Z. Wang, and Y. Liu, "Attribute-controlled face photo synthesis from simple line drawing," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2017, pp. 2946–2950.
[7] C. Galea and R. A. Farrugia, "Forensic face photo-sketch recognition using a deep learning-based architecture," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1586–1590, 2017.
[8] Z. Deng, X. Peng, Z. Li, and Y. Qiao, "Mutual component convolutional neural networks for heterogeneous face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3102–3114, 2019.
[9] Y. Jin, J. Lu, and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 640–652, 2015.
[10] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity-Based Pattern Recognit.*, 2015, pp. 84–92.
[11] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.
[12] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition,"
[13] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737
[14] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," 2014, *arXiv:1412.7755*. [Online]. Available: http://arxiv.org/abs/1412.7755
[15] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9215–9223.
[16] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.
[17] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[18] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "ABCNN: Attention-based convolutional neural network for modeling sentence pairs," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 259–272, 2016.
[19] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[20] J. Hu, L. Shen, and G. Sun, ‘‘Squeeze-and-excitation networks,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, ‘‘CBAM: Convolutional block attention module,’’ in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[22] M. D. Zeiler and R. Fergus, ‘‘Stochastic pooling for regularization of deep convolutional neural networks,’’ in *Proc. 1st Int. Conf. Learn. Represent. (ICLR)*, 2013.

[23] S. Prabhu and I. Pe'Er, ‘‘Overlapping pools for high-throughput targeted resequencing,’’ *Genome Res.*, vol. 19, no. 7, pp. 1254–1261, 2009.

[24] K. He, X. Zhang, S. Ren, and J. Sun, ‘‘Spatial pyramid pooling in deep convolutional networks for visual recognition,’’ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.

[25] J. Xiang and G. Zhu, ‘‘Joint face detection and facial expression recognition with MTCNN,’’ in *Proc. 4th Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, 2017, pp. 424–427.

[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, ‘‘Learning deep features for discriminative localization,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[27] C. Peng, N. Wang, J. Li, and X. Gao, ‘‘DLFace: Deep local descriptor for cross-modality face recognition,’’ *Pattern Recognit.*, vol. 90, pp. 161–171, 2019.

[28] C. Galea and R. A. Farrugia, ‘‘Matching software-generated sketches to face photographs with a very deep CNN, morphed faces, and transfer learning,’’ *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 6, pp. 1421–1431, 2018.

[29] P. Mittal, M. Vatsa, and R. Singh, ‘‘Composite sketch recognition via deep network—A transfer learning approach,’’ in *Proc. Int. Conf. Biometrics (ICB)*, 2015, pp. 251–256.

[30] X. Wang and X. Tang, ‘‘Face photo-sketch synthesis and recognition,’’ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, 2008.

[31] W. Wan, Y. Gao, and H. J. Lee, ‘‘Transfer deep feature learning for face sketch recognition,’’ *Neural Comput. Appl.*, vol. 31, no. 12, pp. 9175–9184, 2019.

LIANG FAN received the master's degree in software engineering from Inner Mongolia Agricultural University. She is currently pursuing the Ph.D. degree with the School of Computer Science and Informatics, Cardiff University.

XIANFANG SUN received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 1994. He is currently a Senior Lecturer with the School of Computer Science and Informatics, Cardiff University, U.K. His main research interests include computer vision, computer graphics, pattern recognition, and artificial intelligence.

PAUL L. ROSIN received the B.Sc. degree in computer science and microprocessor systems from the University of Strathclyde, Glasgow, in 1984, and the Ph.D. degree in information engineering from the City, University of London, in 1988. He is currently a Full Professor with the School of Computer Science and Informatics, Cardiff University. His main research interests include non-photorealistic rendering, mesh processing, and computer vision.