International Journal on Document Analysis and Recognition (IJDAR) (2021) 24:33–48
https://doi.org/10.1007/s10032-021-00364-6

SPECIAL ISSUE PAPER

Cross-modal photo-caricature face recognition based on dynamic multi-task learning

Zuheng Ming[1] · **Jean-Christophe Burie**[1] · **Muhammad Muzzamil Luqman**[1]

Received: 17 February 2020 / Revised: 25 December 2020 / Accepted: 24 February 2021 / Published online: 16 March 2021

Abstract
Face recognition of realistic visual images (*e.g.,* photos) has been well studied and made significant progress in the recent decade. However, face recognition between realistic visual images/photos and caricatures is still a challenging problem. Unlike the photos, the different artistic styles of caricatures introduce extreme non-rigid distortions of caricatures. The great representational gap between the different modalities of photos and caricatures is a big challenge for photo-caricature face recognition. In this paper, we propose to conduct cross-modal photo-caricature face recognition via multi-task learning, which can learn the features of different modalities with different tasks. Instead of manually setting the task weights as in conventional multi-task learning, this work proposes a dynamic weights learning module which can automatically generate/learn task weights according to the training importance of tasks. The learned task weights enable the network to focus on training the hard tasks instead of being stuck in the overtraining of easy tasks. The experimental results demonstrate the effectiveness of the proposed dynamic multi-task learning for cross-modal photo-caricature face recognition. The performance on the datasets CaVI and WebCaricature show the superiority over the state-of-art methods. The implementation code is provided here. (https://github.com/hengxyz/cari-visual-recognition-via-multitask-learning.git).

Keywords Photo-caricature face recognition · Dynamic multi-task learning · Deep CNNs

1 Introduction

In the past decade, face recognition with realistic visual images has advanced considerably. Face recognition based on the powerful representation learning with the deep Conventional Neural Networks (CNNs) [1–4] has achieved beyond human performance on some standard benchmarks such as LFW [5] and YTF [6]. Rather than the conventional methods based on the hand-craft features, *e.g.,* LBP [7], Gabor-LBP [8], HOG [9] and SIFT [10], the deep learning-based methods mitigate the problems of occlusion, illumination variation or large pose by leveraging the enormous data for learning better generalized feature representations for face recognition.

✉ Zuheng Ming
zuheng.ming@univ-lr.fr

Jean-Christophe Burie
jcburie@univ-lr.fr

Muhammad Muzzamil Luqman
mluqma01@univ-lr.fr

[1] Laboratory L3i, La Rochelle University, 17402 La Rochelle, France

Nonetheless, the non-rigid variation in facial expressions or the non-rigid distortion in the caricatures (as shown in Fig. 1) is still a challenge for face recognition. Unlike the realistic visual face images (*e.g.,* photos), caricatures are artistic drawings with exaggerations to strengthen certain instinct features. The diverse artistic styles of caricature introduce large variations not only between photos and caricatures but also between different caricatures with the same identity [11]. Similar to photo-sketch face recognition [12,13], photo-caricature face recognition is one kind of heterogeneous face recognition scenes which is challenging and important to the understanding of face perception [14].

As face recognition focusing on multi-modal facial images such as near-infrared [15,16], forensic sketches [12,13], depth imagery [17,18], etc., photo-caricature face recognition is also a cross-modal problem where one of the modality is a caricature [19].

Thanks to the inherent similarity in the structure of a face captured using different modalities, transfer learning has been successfully applied to directly match a photo to a face capture using different modalities [12,13,15,16]. The transfer learning-based methods pretrain deep networks (*e.g.,* deep

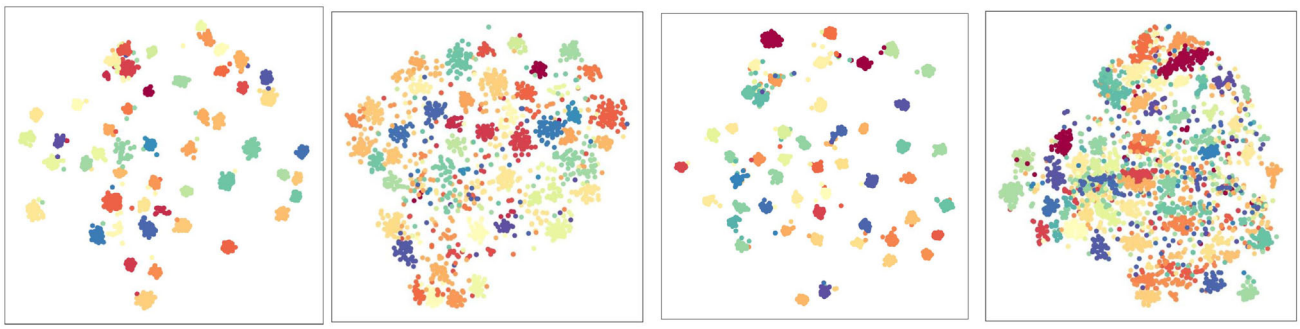
Fig. 1 Photos and caricatures of Bill Clinton and Barack Obama from the datasets CaVI and WebCaricature, respectively. The different artistic styles result in the extreme distortion of caricatures, which leads a large variation between caricatures even if with the same identity

CNNs) using a large face database of photos and then finedtune the pretrained deep networks on the small database of a specific modality. However, photo-caricature face recognition is a challenging task due to the extreme levels of distortions of caricature [19].

WebCaricature [11] firstly proposed to directly adopt pretrained VGG-Face [2] to extract features of photos and caricatures with same process. Then the extracted features are fed into a classifier trained by traditional metric learning methods such as principal component analysis (PCA), kernel discriminant analysis (KDA) [20] etc. to decide whether the photos and caricatures are from the same person. Whereas, the performance of WebCaricature is limited using the model pretrained on photos to extract the feature of caricatures without considering the different modalities of photos and caricatures. Figure 2 shows the representational gap of different modalities of photos and caricatures, which suggests that photo-caricature face recognition with a wide divergence between the two modalities is a not a trivial task.

Learning a general representation that can capture the features of different modalities of caricature and photos is crucial for cross-modal photo-caricature face recognition. Rather than the single-task methods such as WebCaricature [11] can only learn a single-modality feature space based on pretrained VGG-Face [2], the multi-task learning-based approaches enable to learn the common representation space of different modalities by training different modality-specific tasks simultaneously [19].

Comparing to single-task learning, multi-task learning has several advantages. Firstly, multi-task learning can employ implicit data augmentation using the different data for training the different tasks simultaneously, which enables the model to learn a more general representation through averaging the different noise patterns of different data [22]. In this work, we train a multi-task network to perform photo-caricature face verification, caricature identification and photo identification simultaneously using both the caricature and photos as training data. Since the caricatures and photos have different noise patterns, multi-task learning enables the network to learn a more general feature by ignoring the data-dependant noise. Instead, single-task learning bears the risk of overfitting the noise of the task which impedes the network to learn a general feature. Secondly, multi-task learning can also help the model to learn the difficult features. Some features are difficult to learn for some tasks, while being easy to learn for another task. In this work, we can see in the following sections that the features of caricatures are more difficult than the features of photos. By integrating the different tasks, such as combining caricature identification task with photo-caricature face recognition task



(a) V2V classification **(b)** V2C classification **(c)** C2C classification **(d)** C2V classification

Fig. 2 The representational gap of learned features of different modalities, *i.e.,* photos and caricatures. The photos and caricatures are from the dataset CaVI [19]. The different colors denote the different 50 identities. **a** V2V classification: training a deep CNNs on training photos and identifying test photos using the trained CNNs (98.10% of identifying accuracy); **b** V2C classification: identifying test caricatures using the pretrained CNNs in (a) (53.60% of identifying accuracy); **c** C2C classification: training a deep CNNs on training caricatures and identifying test caricatures using the trained CNNs (78.20% of identifying accuracy); **d** C2V classification: identifying test photos using the pretrained CNNs in (c) (41.80% of identifying accuracy). The deep CNNs used in all cases have the same architecture. The visualisation is implemented by t-SNE [21]

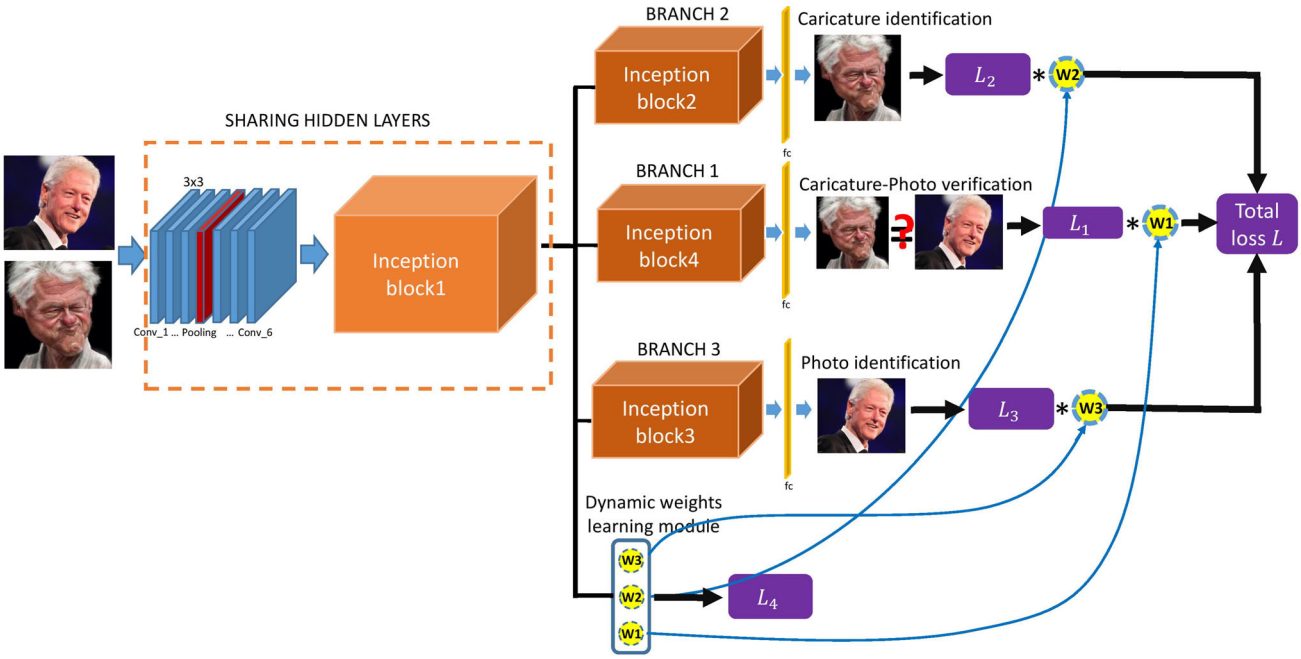


Fig. 3 The proposed dynamic multi-task learning network for cross-modal photo-caricature face recognition. Different recognition modalities are learned by the different tasks. The dynamic weights learning module connecting to the end of sharing hidden layers can update the task weights according to the importance of tasks during the training of network

into multi-task learning network, the model can learn the features of caricatures when either training the caricature identification task or the photo-caricature face recognition task.

There are mainly two types of multi-task learning methods: Hard parameters sharing-based methods and soft parameters sharing-based methods. In the context of deep learning, multi-task learning is typically done with either hard or soft parameter sharing of hidden layers [22]. Hard parameter sharing are the most commonly used approaches to multi-task learning, in which the first several hidden layers are shared between all tasks, while keeping several task-specific output layers for each task [23–28]. In other hand, in soft parameter sharing-based multi-task learning methods, each task has its own model with its own layers. There is no sharing layers between the tasks [29–31]. Since caricatures and photos having different modalities still share some common intrinsic facial features such as the facial topological structure etc., the hard parameter sharing-based multi-task learning method is adopted in this work instead of using the soft parameters sharing architecture as Siamese couple networks [19].

Multi-task learning is substantially an optimization problem for multiple objectives. The different tasks may have different importance or training difficulty, therefor, how to find the optimal weights of tasks is an important issue in multi-task learning. Many works prove that the performance varies in function of the weights in multi-task learning, and the optimal performance can be obtained by the weighted task with different weights [27]. Thus it is unwise to assign equal weights of tasks for multi-task learning as described in [32].

There are mainly two ways to search the optimal weights for multi-task learning: (1) the static methods; (2) the dynamic methods. In the static methods, the weights of tasks are searched either manually by experimental methods such as [19,24] or by a greedy search [25]. The found optimal weights are assigned to the tasks and fixed during the training. Searching task weights manually is laborious and low efficiency, while the greedy search method is time consuming. Rather than the static methods, the dynamic methods enable to adapt the task weights automatically according to the variation of the losses, the gradients, the uncertainty of tasks and so on [26,27,30,33]. However, these methods all introduce the hyperparameters for learning the task weights. Yin et al. [28] updates the dynamic weights of tasks with the total loss of the network. Nevertheless, this method results in the network being stuck in the overtraining of the easy task and the undertraining of the hard task.

In this work, we propose a multi-task learning network with a dynamic weights learning module based on the deep CNNs for photo-caricature face recognition (see Fig. 3). Instead of simply updating the dynamic weights with the total loss of network as in [28], we propose a dynamic weights learning module connecting to the sharing layers of network

to learn the task weights (see Fig. 3). With a designed loss function based on the importance of tasks, the learned task weights from the dynamic weights learning module enable the network to focus on training the hard task instead of being stuck in the overtraining of easy task.

Inspired by Yin et al. [28] using cross-entropy loss to update task weights and Nguyen et al. [34] using conditional cross-entropy to measure the hardness of task being transferred to another task in transfer learning, we propose to use the cross-entropy loss of each task to measure the importance/hardness of tasks during the training. In this work, the cross-entropy loss of different tasks either represents the error of the photo-caricature face recognition, or the error of photo or caricature identification. The higher cross-entropy loss of task indicates that the task is harder to be trained during the training of network. Instead, the task with lower cross-entropy loss is easier to be trained. In order to focus on training the hard task, the task with the higher loss being more important during the training should be assigned a larger task weight. Moreover, no hyperparameters are introduced for learning the task weights in the proposed dynamic weights learning module.

As shown in Fig. 3, three different tasks (*i.e.,* photo-caricature face verification, caricature identification, and photo identification), with three different branches connecting to the sharing hidden layers, are integrated into the proposed multi-task learning network. Each output of the dynamic weights learning module serves as a dynamic task weight of a task.

In summary, the main contributions of this paper are:

- We propose a multi-task learning approach with dynamic weights for cross-modal photo-caricature face recognition, which can model the different recognition modalities by the different tasks.
- The proposed dynamic weights learning module without introducing additional hyperparameters can lead multi-task learning to train the hard task primarily instead of the overtraining of the easy task, which results multi-task learning more efficiently.
- Both the theoretical analysis and the experimental results demonstrate the effectiveness of the proposed method for updating the dynamic weights of tasks during the training.
- We have demonstrated that, for all the three recognition tasks, the proposed multi-task learning can outperform the state-of-the-art performance on the datasets CaVI and WebCaricature.

The remainder of this paper is organized as follows: Sect. 2 briefly reviews the related works; Sect. 3 presents the approach of multi-task learning with dynamic weights. Section 4 describes the architecture of the dynamic multi-task network proposed in this work and Sect. 5 shows the experimental results. Finally, in Sect. 6, we draw the conclusions and present the future works.

2 Related works

2.1 Cross-modal face recognition

(1) Photo-caricatureface recognition Before the representation learning by the deep CNNs, the handcrafted features such as the local descriptors HOG, LBP, Gabor-LBP, SIFT and Fisher Vector have been widely used for face recognition [7]. By virtue of the deep neural networks especially the deep CNNs, face recognition has made a series of breakthrough in the recent decade. DeepFace [1] firstly introduces a siamese network architecture for the face verification and has achieved 97.35% on the LFW and 91.4% on the YTF. DeepID [35] series using more than 200 CNNs for face verification to gain a better performance (99.15% on LFW). FaceNet [3] proposes triplet loss to learn embedding features for face recognition and achieve the state-of-art on LFW (99.63%) and YTF (95.12%). VGG face [36] continues to implement the triplet loss on the VGG net. Wen and al. [37] propose the center loss joint with softmax to achieve the state-of-the-art performance. Recently SphereFace [4] proposes a revised softmax to learn angularly discriminative features and achieves the state-art-art performance on dataset MegaFace [38].

Due to the challenge of the cross-modal heterogeneous face matching problem and also the lack of the dataset, the photo-caricature face recognition is not sufficiently studied especially with the deep learning based methods. Huo and al. [11] propose a large caricature dataset WebCaricature consisting of 252 people with 6024 caricatures and 5974 photos. A baseline for caricature face verification and identification is also proposed, respectively. It shows that the performance of the deep learning based method with pretrained VGG-Face is significant better than the hand-craft feature based methods such as SIFT, Gabor etc. However, the performance of the proposed method is still limited that the best performance for photo-caricature face verification is 57.22% of validation rate (recall rate) % @ FAR 1%. Meanwhile, it achieves 55.41% @ Rank-1 accuracy for caricature to real visual image identification and 55.53% @ Rank-1 accuracy for real visual image to caricature identification. Garg et al. [19] propose a CNN-based coupled-networks CaVINet consisting of couple of 13 convolutional layers of VGGFace for cross-model caricature-verification and caricature identification. Besides, this work also introduce a new publicly available dataset (CaVI) that contains caricatures and visual images of 205 identities, which has 5091 caricatures and 6427 visual images. The CaVINet can achieve 91.06% accuracy

for the photo-caricature face verification task, 85.09% accuracy for caricature identification task and 94.50% accuracy for caricature identification task. It notes that the weights of tasks are manually searched by the experimental method.

(2) Photo-sketch face recognition Similar to photo-caricature face recognition, the photo-sketch face recognition is another type of cross-modal face recognition. As well as photo-caricature face recognition, transfer learning is also widely used to directly match photos to sketches. The pretrained deep networks using a large face database of photos are fine-tuned using small sketch database for photo-sketch face recognition [12,13]. Instead of directly match photos to sketches, Zhang et al. [39] proposed to use a fully convolutional network to transform the photos to sketches and then compared the transformed sketches with the target sketches for conducting photo-sketch face recognition. As the Generative Adversarial Network (GAN) shows powerful ability for generating high-quality photos from sketches and sketches from photos [40], Wang et al. [41] proposed a multi-adversarial network to transform photos to sketches or sketches to photos for photo-sketch face recognition.

There are also other types of cross-modal face recognition such as face recognition between Near-Infrared Spectrum (NIS) images and Visible Light Spectrum (VIS) images [42–44]. However, these topics are beyond the scope of this paper.

2.2 Multi-task learning

Multi-task learning has been used successfully across many areas of machine learning [22], from natural language processing and speech recognition [45,46] to computer vision [23]. Fast R-CNN [23] uses a multi-task loss to jointly train the classification and bounding-box regression for object detection. The classification task is set as the main task with the weight 1 and the bounding-box regression is set as the side task weighted by λ. The author also shows the improvement of multi-task learning for object detection comparing to multi-task learning. Hyperface [24] proposed a multi-task learning algorithm with static weights for face detection, landmarks localization, pose estimation and gender recognition using deep CNNs. Tian et al. [25] fix the weight for the main task to 1, and obtain the weights of all side tasks via a greedy search within 0 and 1. In [26] the weights is updated dynamically by the loss of the gradients meanwhile an hyperparameter is introduced for balancing the training of different tasks. [27] introduces a uncertainty coefficient θ to revise the loss function which can be fixed manually or learned based on the total loss. Zhang et al. [33] introduce an hyperparameter ρ as a scale factor to calculate the dynamic weight λ_t of face attributes recognition. Yin et al. [28] proposed a multi-task model for face pose-invariant recognition in which the main task is face identification and the side tasks are the classification of face pose, facial expressions and face illumination. The weight of main task is set 1 and the weights of the side tasks are assigned by the dynamic weights generated by the softmax layer. Since the dynamic weights of tasks are updated by the total loss of network, the training of multi-task learning is stuck in the overtraining of the easy task while the hard task is undertraining.

3 Architecture of dynamic multi-task learning network

Although caricatures and photos have different recognition modalities but they still share some common features of face patterns, such as the similar topological structures of the eyes, nose, mouth, etc. Thus we wish the network can also learn the common features between different modalities when learning the modality-specific features. In this work, we propose to use deep multi-task learning network with the hard parameter sharing structure [22] to conduct the cross-modal photo-caricature face recognition, in which the different tasks share the first several hidden layers to capture the modality-common features between all tasks (see Fig. 3). In this work, we simplify the network to use one-stem-based network as the hidden sharing layers to learn the common features across modalities, and enforce the task-specific branches to learn the modality-specific features.

As shown in Fig. 4, the deep neural network is constructed by the Inception-ResNet [47] blocks. The three branches are, respectively, dedicated to caricature identification, photo identification and photo-caricature face verification. The three branches have almost identical structures to facilitate the initialization of branches from the pretrained branch. Specifically, BRANCH 1 can extract the feature embeddings from the bottleneck layer for photo-caricature face verification, and BRANCH 2 and 3 use the fully connected softmax layer to calculate the probabilities of the possible identities of the input caricatures or photos. The details of the architecture of the proposed multi-task learning network are shown in Table 1.

Dynamic weights learning module The weights of tasks are generated automatically by the dynamic weights learning module during the training (see Fig. 3). The dynamic weights learning module consists of a fully-connected layer with softmax normalization (*i.e.,* the so-called softmax layer), which is connected to the end of the sharing hidden layers as shown in Fig. 4. Since the task weights are used to weight the importance of different tasks during the training, we connect the dynamic weights learning module to the last sharing hidden layer which can leverage the extracted common information between tasks to learn the task weights. The dynamic weights learning module has three outputs corresponding to the three tasks integrated with the network. Each output of dynamic weights learning module serves as a weight of a

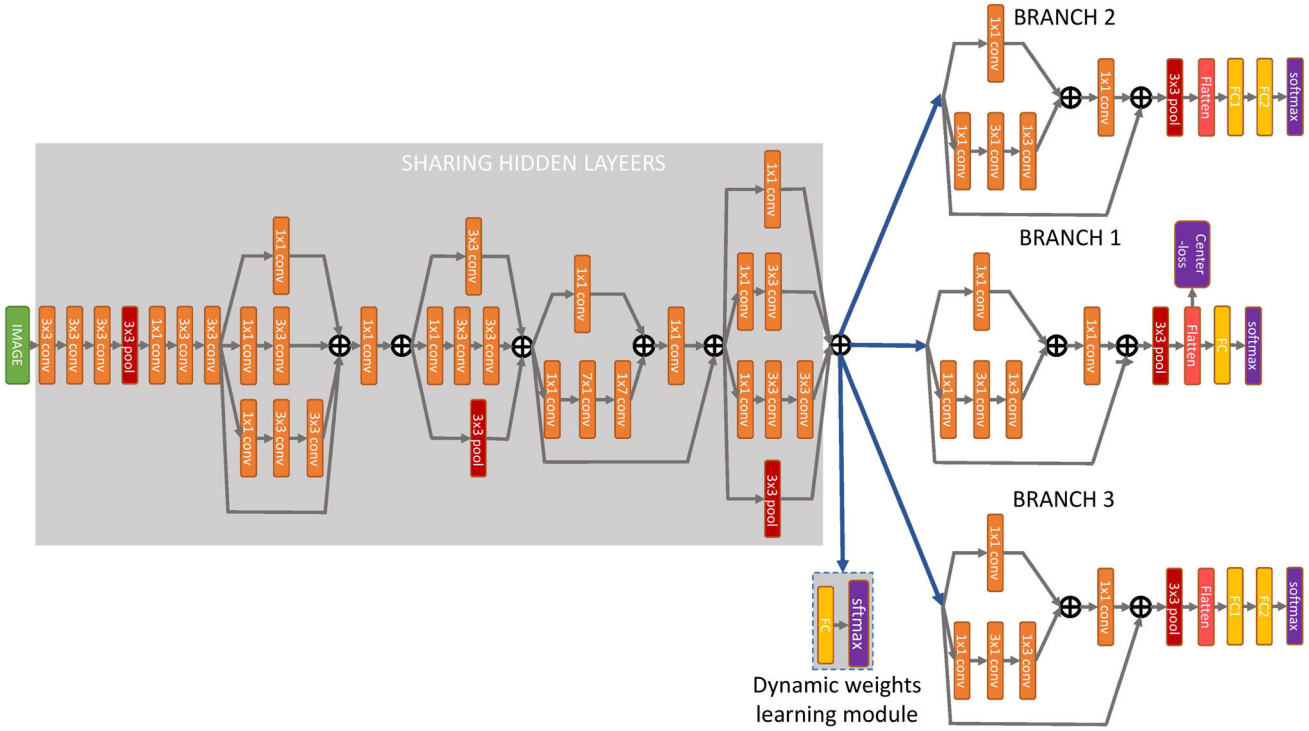


Fig. 4 The architecture of the proposed dynamic multi-task learning network for cross-modal photo-caricature recognition. The weights of tasks can be automatically learned/generated by the dynamic weights learning module. The network is based on the Inception-ResNet structure

task $w_i \in \{w_1, w_2, w_3\}$. Since the task weights are the outputs of softmax layer, the sum of the task weights w_i equals to one, *i.e.,* $\sum_{i=1}^{3}(w_i) = 1$. The dynamic task weights are determined by the training importance/hardness of tasks, thus the loss function $\mathcal{L}_4$ for optimizing the parameters of the dynamic weights learning module is in function of the losses of tasks which represent the importance/hardness of tasks during the training. More details about the loss function $\mathcal{L}_4$ are described in Sect. 4.

Note that the dynamic weights learning module and the dynamic task weights are only used for training the network. During the inference/test phase, the task weights and the dynamic weights learning module are no longer used.

Prediction of the identity of caricature/photo: When the network finishes the training, each branch of network can independently conduct the specific task such as photo-caricature face verification, caricature identification and photo identification. Given a caricature or a photo as the input image of the network, the probabilities of all possible identifies of the given image are predicted as the output of BRANCH 2 (caricature identification) or BRANCH 3 (photo identification). The identity with the highest probability is the predicted identity of the given photo or caricature. For photo-caricature face verification, it needs to input a pair of images, *i.e.,* sequentially inputting a photo and a caricature into the network. Then the embeddings being the feature of the input image are extracted successively using the bottleneck layer of BRANCH 1. As the widely used protocol for face verification [3,35,37,48], photo-caricature face verification here is conducted by thresholding the $l2$-norm distance between the embeddings of the given photo and caricature. If the $l2$-norm distance between the embeddings is smaller than the predefined threshold, the given photo and caricature are considered having the identity, vice versa. The threshold is normally obtained from the training/validation dataset. Note that, when conducting photo-caricature face verification given the input pair of images, caricature identification and photo identification can be also performing simultaneously with BRANCH2 and BRANCH3.

4 Multi-task learning with dynamic weights

Multi-task learning is an optimization problem for multiple objectives. The conventional multi-task learning [23–25] often sums up the different tasks with the fixed task weights:

$$\mathcal{L}(\mathbf{X}; \Theta) = \sum_{i=1}^{T} w_i \mathcal{L}_i(\mathbf{X}; \Theta_i) \tag{1}$$

where T is the number of tasks, X is the input of the model and $\Theta = \{\Theta_i\}$ are the parameters of the model.

Table 1 Details of the architecture of the proposed dynamic multi-task learning network

Layer	Kernel	#1x1	#3x3	#3x1	#1x3	#1x7	#7x1	BRANCH
conv1	3x3x32,2							
conv2	3x3x32,1							
conv3	3x3x64,1							
maxpool1	3x3,2							
conv4	1x1x80,1							
conv5	3x3x192,1							
conv6	3x3x256,2							
Inception(1a)		32,1						
Inception(1b)		32,1	32,1					
Inception(1c)		32,1	32,1 (2)					
conv7	1x1x192,1							
Inception(2a)		384,2						
Inception(2b)		192,1	192,1; 256,1					
Inception(2c)	maxpool 3x3,2							
Inception(3a)		128,1						
Inception(3b)		128,1				128,1	128,1	
Inception(4a)		256,1						
Inception(4b)		256,1	256,2					
Inception(4c)		256,1	256,1; 256,2					
Inception(4d)	maxpool 3x3,2							
Inception(5a)		192,1						1
Inception(5b)		192,1		192,1	192,1			1
conv8	1x1x192,1							1
avgpool1								1
fullyconn1								1
Inception(7a)		192,1						2
Inception(7b)		192,1		192,1	192,1			2
conv9	1x1x192,1							2
avgpool2								2
fullyconn3								2
fullyconn4								2
Inception(9a)		192,1						3
Inception(9b)		192,1		192,1	192,1			3
conv10	1x1x192,1							3
avgpool3								3
fullyconn4								3
fullyconn5								3

The kernel is specified as rows x cols x depth, stride. The repeat number of the kernel is denoted in the bracket. BRANCH denotes the branch to which the block belongs. #1x1,#3x3,.., denote the conv kernel used in the inception block

$\{w_i\}$ are the weights of tasks which are invariable during the training of the model. Comparing to single-task learning, multi-task learning can gain a better performance by joint learning different tasks with the appropriate task weights as shown in [27]. However, manually searching the optimal task weights is laborious and time-consuming. Here, we propose an approach based on deep multi-task learning with dynamic task weights, which can learn and update the task weights according to the training hardness/importance of tasks during the training of the network.

(I) **Dynamic multi-task learning**: Instead of using fixed task weights, dynamic multi-task learning combines different tasks with the dynamic task weights during the training processing. The total loss $\mathcal{L}$ of dynamic multi-task learning is defined as follows:

$$\mathcal{L}(\mathbf{X};\Theta;\Psi)=\sum_{i=1}^{T}w_i(\Psi)\mathcal{L}_i(\mathbf{X};\Theta_i) \tag{2}$$

where T is the number of tasks, here $T=3$. X and $\Theta=\{\Theta_i\}$ are the input and parameters of the network. $\{\alpha_i=w_i(\Psi)\}$ are the dynamic task weights which are learned by the dynamic weights learning module (see Fig. 3). Ψ are the parameters of the dynamic weights learning module. Since the dynamic weights learning module consists of a softmax layer (*i.e.,* a fully-connected layer with softmax function) as shown in Fig. 4, Ψ are actually the parameters of the fully-connected layer in the dynamic weights learning module. Note that $\Psi\not\subset\Theta$, so the parameters of the dynamic weights learning module Ψ and the parameters of the network Θ are optimized, respectively, by their own loss functions. Particularly, since the dynamic task weights $\{\alpha_i\}$ are the outputs of the softmax layer, $\sum\alpha_i=1$. When $\alpha_i=1, i\in[1,2,3]$, multi-task learning is degraded as single-task learning of task i.

Training protocol: Since the parameters of the dynamic weights learning module Ψ and the parameters of the network Θ are independent, the entire training of the dynamic multi-task learning network includes two parts: the optimization of the parameters of the network Θ using the total loss $\mathcal{L}$ from Equation 2 and the optimization of the parameters of the dynamic weights learning module Ψ using its own loss $\mathcal{L}_4$ from Equation 12. The optimization of Ψ and Θ can be conducted simultaneously in a parallel way.

$$\Theta^{t-1}-\eta\frac{\partial\mathcal{L}(\Theta)}{\partial\Theta}\mapsto\Theta^{t} \tag{3}$$

$$\Psi^{t-1}-\eta\frac{\partial\mathcal{L}_4(\Psi)}{\partial\Psi}\mapsto\Psi^{t} \tag{4}$$

where $\eta\in(0,1)$ is the learning rate.

(II) **Photo-caricature face verification task loss** $\mathcal{L}_1$: The loss for photo-caricature face verification task is measured by the center loss [37] joint with the cross-entropy loss of BRANCH 1. The loss function $\mathcal{L}_1$ is given by:

$$\mathcal{L}_1(\mathbf{X};\Theta_1)=\mathcal{L}_{s1}(\mathbf{X};\Theta_1)+\beta\mathcal{L}_c(\mathbf{X};\Theta_1) \tag{5}$$

where $\mathcal{L}_{s1}$ is the cross-entropy loss and $\mathcal{L}_c$ is the center loss of BRANCH 1. Θ_1 are the parameters of the network corresponding to the task 1, which include the parameters of the sharing hidden layers and the BRANCH 1. β is a hyperparameter to weight the center loss. The weighted center loss $\mathcal{L}_c$ can be treated as a regularization item of the softmax loss $\mathcal{L}_{s1}$. The softmax loss $\mathcal{L}_{s1}$ is given by:

$$\mathcal{L}_{s1}(\mathbf{X};\Theta_1)=\sum_{k=1}^{K}-y_k log\,P(y_k=1|\mathbf{X},\Theta_1) \tag{6}$$

where K is the number of identities of the input images $\mathbf{X}$, $y_k\in\{0,1\}$ is the binary indicator function if the class label c is the correct classification for the input image X, $P(y_k=1|\mathbf{X},\Theta_1)$ is the predicted probability of class c of the given image X. The center loss $\mathcal{L}_c$ is given by:

$$\mathcal{L}_c(\mathbf{X};\Theta_1)=||f^{\Theta_1}(\mathbf{X})-C_{y_k}|| \tag{7}$$

Where the C_{y_k} is the center of the class to which $\mathbf{X}$ belonging, $f^{\Theta_1}(\mathbf{X})$ are feature embeddings of $\mathbf{X}$ extracted from the bottle layer of BRANCH 1.

(III) **Caricature identification task loss** $\mathcal{L}_2$: The loss function $\mathcal{L}_2$ is the cross-entropy loss of BRANCH 2, which is given by:

$$\mathcal{L}_2(\mathbf{X};\Theta_2)=\sum_{k=1}^{K}-y_k log\,P(y_k=1|F_p(\mathbf{X}),\Theta_2) \tag{8}$$

where $F_p(\mathbf{X})$ is a function to filter the photos in the input images $\mathbf{X}$, since the caricature identification task only needs caricatures for the training. K is the number of identities of caricatures filtered from the input $\mathbf{X}$. Θ_2 are the parameters corresponding to the task 2, which include the parameters of the sharing hidden layers and the BRANCH 2. As well as task 1, $y_k\in\{0,1\}$ is the binary indicator function and $P(y_k=1|F_p(\mathbf{X}),\Theta_2)$ is the predicted probability of the correct class of the given caricature.

(IV) **Photo identification task loss** $\mathcal{L}_3$: The $\mathcal{L}_3$ is the cross-entropy loss of BRANCH 3 given by:

$$\mathcal{L}_3(\mathbf{X};\Theta_3)=\sum_{k=1}^{K}-y_k log\,P(y_k=1|F_c(\mathbf{X}),\Theta_3) \tag{9}$$

where $F_c(\mathbf{X})$ is a function to filter the caricatures in the input images $\mathbf{X}$ for the photo identification task. K is the number of identities of photos filtered from the input $\mathbf{X}$. Θ_3 are the parameters of the sharing hidden layers and the BRANCH 3. Similarly, $y_k\in\{0,1\}$ is the binary indicator function and $P(y_k=1|F_c(\mathbf{X}),\Theta_3)$ is the predicted probability of the correct class of the given photo.

(V) **Generation of the dynamic weights**: The dynamic weights $\{\alpha_i=w_i(\Psi)\}$ are generated by the softmax layer in the dynamic weights learning module (*i.e.,* the dynamic weights are actually the outputs of the softmax layer):

$$w_i(\Psi)=\frac{e^{f^{\Psi_i}(\mathbf{Z})}}{\sum_j^T e^{f^{\Psi_j}(\mathbf{Z})}} \tag{10}$$

where $\mathbf{Z}$ is a vector obtained by flattening the last sharing hidden layer, which is the input of the dynamic weights learning module (see Fig. 4). Ψ are the parameters of the

fully-connected layer in the dynamic weights learning module, in which $\Psi = \{\psi_j\}_{j=1}^{T}$. T is the number of tasks, here T=3. f^{ψ_i} is activation function of the fully-connected layer which is given by:

$$f^{\psi_i}(\mathbf{Z}) = \psi_i \mathbf{Z}^T + b_i \tag{11}$$

Note that, here we only use the linear mapping function rather than the widely used nonlinear ReLU activation function. Since ReLU activation function discards the values less than zero, it would shrink the range of the generated dynamic weights.

(VI) **Dynamic weights learning loss** $\mathcal{L}_4$: We propose a new loss function of dynamic weight learning module to learn the dynamic weights which can drive the network to focus on training the hard task:

$$\mathcal{L}_4(\Psi) = \sum_{i=1}^{T} \frac{w_i(\Psi)}{\mathcal{L}_i} \quad s.t. \quad \sum w_i = 1 \tag{12}$$

where Ψ are the parameters of the dynamic weights learning module, $w_i(\Psi)$ are the learned dynamic weights and $\mathcal{L}_i$ are the loss of the task i ($i \in [1, .., T]$). Note that, $\mathcal{L}_i$ is in function of the parameters of the network Θ_i which are independent with Ψ. Thus, $\mathcal{L}_i$ is a constant when using the dynamic weights learning loss $\mathcal{L}_4$ to optimize the parameters of dynamic weights learning module Ψ.

We can see that $\mathcal{L}_4$ can prevent the network from over-training the easy task by generating smaller task weight for the easier task. For example, after some training epochs, the loss of the task i begins to decrease showing that the task i becomes easier for training. However, when the loss $\mathcal{L}_i$ becomes smaller, the loss of the dynamic weights learning module $\mathcal{L}_4$ would become larger (see Equation 12). Since the optimization goal of the parameters of the dynamic weights learning module is to decrease the loss $\mathcal{L}_4$, the parameters of the dynamic weights learning module Ψ then would be optimized to decrease $w_i(\Psi)$ aiming to decrease the loss $\mathcal{L}_4$. Therefore, when the task become easier with a small loss during the training, the dynamic weights learning module would generate a smaller weight for the task. Consequently, the larger task weights would be generated for the harder task with large loss due to $\sum w_i = 1$.

(VII) **Quantitative Analysis of the dynamic weights** Here we show how the proposed dynamic weights drive the network focus on training the hard task. Considering Eqs. 10 and 11, the gradient of the ψ_i can be given by

$$\nabla\psi_i = \frac{\partial \mathcal{L}_4}{\partial \psi_i} = \frac{1}{\mathcal{L}_i}\frac{\partial w_i(\psi_i)}{\partial \psi_i} = \frac{1}{\mathcal{L}_i}\frac{a_i \sum_{j\neq i}^{T} a_j}{\left(\sum_i^T a_i\right)^2}\mathbf{Z} \tag{13}$$

where $a_i = e^{\psi_i \mathbf{Z}^T + b_i}$, and the update of the parameters is $\psi_i^{t+1} = \psi_i^t - \eta\nabla{\psi_i}^t$ where η is the learning rate. Then the new value of the dynamic weight w_i^{t+1} can be obtained by Eq. 10 with the ψ_i^{t+1}. We assume the $b_i^0 = 0$, $\psi_i^0 = 0$, $\eta = 1$, (this is possible if we initialize the ψ_i, b_i by zero), the ψ_i^t can be given by

$$\psi_i^t = -\sum \frac{1}{\mathcal{L}_i}\frac{a_i \sum_{j\neq i}^{T} a_j}{\left(\sum_i^T a_i\right)^2}\mathbf{Z} \tag{14}$$

if we consider the case for two tasks w_1 and w_2:

$$\begin{aligned}\frac{w_1^t}{w_2^t} &= e^{(\psi_1^t - \psi_2^t)\mathbf{Z}^T} \\ &= e^{\left(\frac{1}{\mathcal{L}_2} - \frac{1}{\mathcal{L}_1}\right)\frac{a_1 a_2}{(a_1+a_2)^2}\mathbf{Z}\mathbf{Z}^T}\end{aligned} \tag{15}$$

We can see that $a_i > 0$ and $ZZ^T \geq 0$, so if $\mathcal{L}_2 < \mathcal{L}_1$ the $\frac{w_1}{w_2} > 1$ namely $w_1 > w_2$. It means if the loss of task1 larger than the loss of task 2, the weight of the task1 is larger than the one of task2. It indicates that the proposed loss function $\mathcal{L}_4$ can well update the weights of tasks to drive the network always train the hard task firstly.

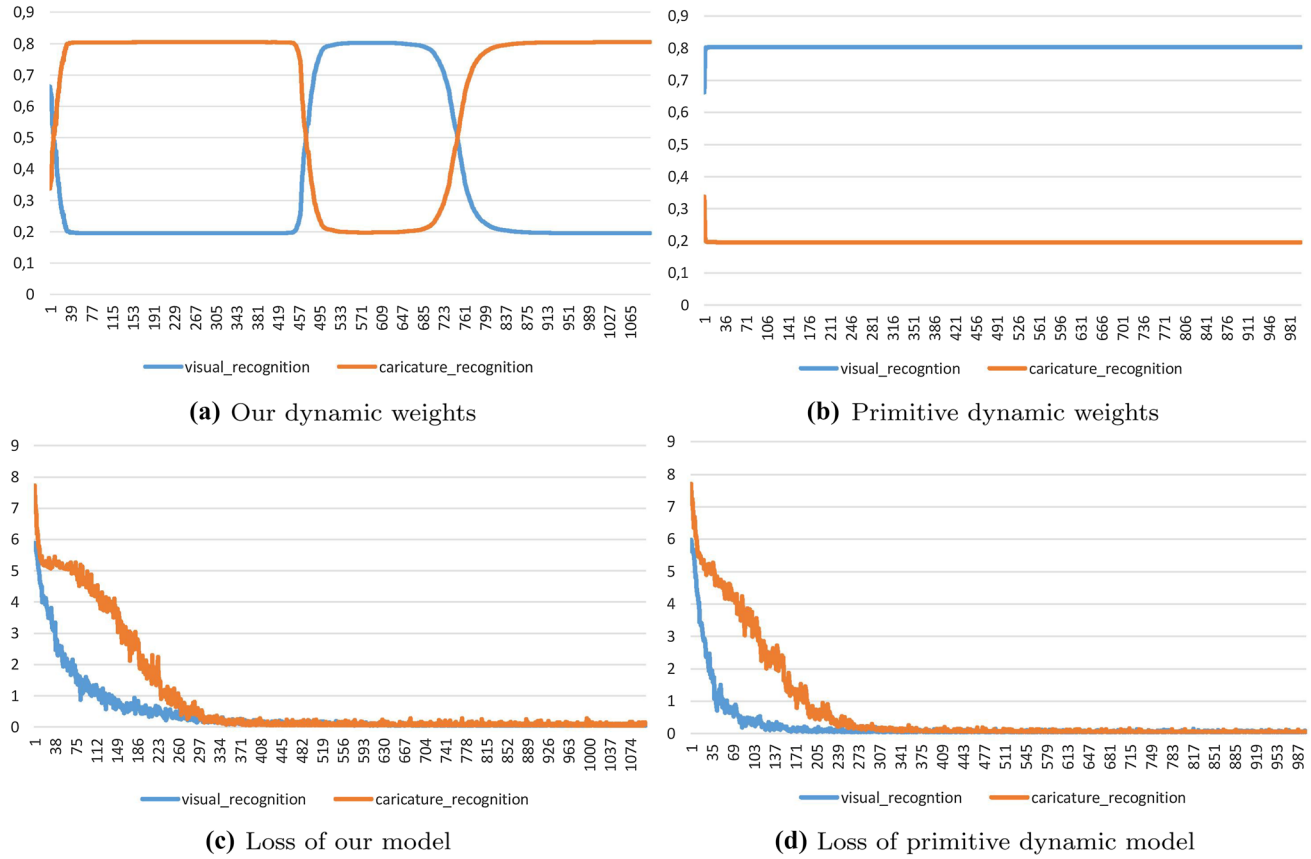
5 Experiments and analysis

5.1 Datasets

CaVI and WebCaricature are so far the largest public datasets for photo-caricature recognition research. In this work, the two datasets are all used to train and evaluate our proposed model.

CaVI contains caricatures and visual images of 205 identities, which has 5091 caricatures ranging from 10-15 images per identity and 6427 visual images (*i.e.,* photos) ranging from 10-15 images per identity. OpenFace [49] is used to extract faces from the scrapped visual images and verify the estimated bounding box manually to ensure the accuracy of the detected faces. The faces are manually extracted for caricatures and only the complete faces were annotated in the dataset.

WebCaricature is a large caricature dataset of 252 people with 6024 caricatures and 5974 photos. For each person, the number of caricatures ranges from 1 to 114 and the number of photos from 7 to 59. The caricatures are labeled manually with 17 landmarks and the landmarks of photos are detected automatically by the software Face++ [50].



(a) Our dynamic weights

(b) Primitive dynamic weights

(c) Loss of our model

(d) Loss of primitive dynamic model

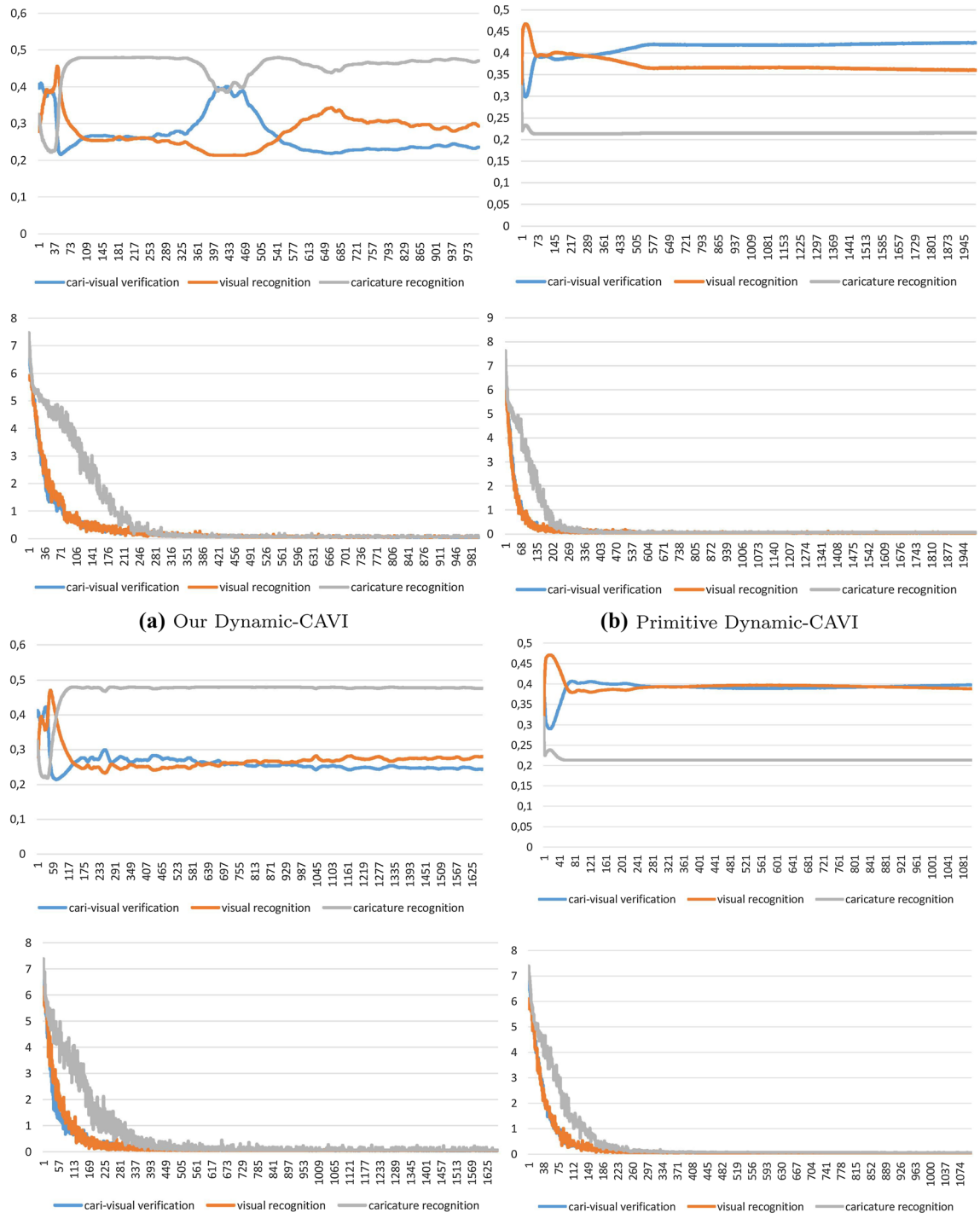
Fig. 5 The comparison of our proposed dynamic multi-task learning method and the primitive dynamic multi-task learning method proposed in [28] performing caricature identification and photo identification on dataset CAVI. The upper row shows the dynamic weights for both methods and the bottom row is corresponding to their losses. (**a**, **c**) are corresponding to our proposed approach and (**b**, **d**) are the primitive dynamic multi-task learning method. The orange curves are corresponding to caricature identification and the blue curves denote photo identification. The horizontal axis is the training iteration (color figure online)

5.2 Pretrained model

Either the dataset CaVI and WebCaricature is relative small to train such a deep CNNs for face recognition. Before the training of the proposed multi-task CNNs, a single-task network consisting of the sharing hidden layers and BRANCH 1 is pretrained for face verification task with large-scale dataset MSCeleb-1M [51]. MTCNN [52] is used to detect the face from the raw images. The RMSprop with the mini-batches of 90 samples is applied for optimizing the parameters. The momentum coefficient is set to 0.99. The learning rate is started from 0.1, and divided by 10 at the 60K, 80K iterations, respectively. The dropout probability is set 0.5 and the weight decay is 5e-5. The network is initialized by Glorot and Bengio [53] with the zero bias. When BRANCH 1 finishes the training as a single-task network, BRANCH 2 and BRANCH 3 can be initialized by the pretrained BRANCH 1.

5.3 Toy example

In order to better demonstrate the effectiveness of the proposed dynamic multi-task learning network for cross-modal face recognition, a toy example with two tasks is conducted on dataset CAVI. Caricature identification and photo identification as two independent tasks are selected to perform the dynamic multi-task learning. Thus, the weight for photo-caricature face verification task w_1 is set to 0 and then $w_2 + w_3 = 1$. In Fig. 5, we compare our dynamic multi-task learning approach with the primitive dynamic multi-task learning method proposed in [28]. We can see that for both methods, caricature identification (denoted in orange) with larger loss is the hard task at the beginning of the training. While our proposed method assigns a larger weight (the orange in (a)) to the hard task enabling the network to focus on training the hard task. Instead, the primitive dynamic method assigns a larger weight to the easy photo identification task with small loss (denoted in blue). In the following training, the dynamic weights generated by our method can adapt



(a) Our Dynamic-CAVI **(b)** Primitive Dynamic-CAVI

(c) Our Dynamic-WebCari **(d)** Primitive Dynamic-WebCari

Fig. 6 The evaluation of our dynamic multi-task learning method and the primitive dynamic multi-task learning method on the datasets CaVI and WebCaricature. In each sub figure, the upper row shows the dynamic weights and the bottom row shows the corresponding loss. The grey curves denote the caricature identification task, the orange curves denote the photo identification task and the blue curves are the photo-caricature face verification task. The horizontal axis is the training iteration (color figure online)

Table 2 The evaluation of different methods for photo-caricature face verification, photo identification and caricature identification (accuracy%) on dataset CaVI

Method	Verification	Photo identification	Caricature identification	V2C	C2V
CaVINet	91.06	94.50	85.09	–	–
CaVINet(TW)	84.32	85.16	86.02	–	–
CaVINet(w/o ortho)	86.01	93.46	80.43	–	–
CaVINet(shared)	88.59	90.56	81.23	–	–
CaVINet(visual)	88.58	92.16	83.36	–	–
Navie Dynamic	93.80	97.60	75.80	61.90	62.80
Ours (Single-verif)	92.46	–	–	–	–
Ours (Single-visual)	–	98.10	-	-	41.80
Ours (Single-cari)	–	–	78.20	53.60	–
Ours (Dynamic MTL)	**94.92**	**98.35**	**85.61**	**80.04**	**64.39**

Best results in each column are in bold

Table 3 The evaluation of different methods for photo-caricature face verification in terms of the validation rate (%) on dataset WebCaricature

Method	VAL@FAR = 0.1%	VAL@FAR = 1%	AUC
SIFT-Land-ITML	5.08±1.82	18.07±4.72	0.841±0.018
VGG-Eye-PCA	21.42±2.02	40.28±2.91	0.896±0.013
VGG-Eye-ITML	18.97±3.90	41.72±5.83	0.911±0.014
VGG-Box-PCA	28.42±2.04	55.53±2.76	0.946±0.009
VGG-Box	34.94±5.06	57.22±6.50	0.954±0.010
Navie Dynamic	38.39±4.58	79.69±1.3	0.961±0.004
Ours (Single-verif)	42.10±3.05	**84.52±0.80**	0.948±0.002
Ours (Dynamic MTL)	**45.82±1.65**	83.20±2.00	**0.987±0.002**

Best results in each column are in bold

Table 4 The evaluation of Caricature to Photo identification (C2P) on dataset WebCaricature

Method	Rank-1 (%)	Rank-10 (%)
SIFT-Land-KCSR	24.87 ± 1.50	61.57 ± 1.37
VGG-Eye-PCA	35.07 ± 1.84	71.64 ± 1.32
VGG-Eye-KCSR	39.76 ± 1.60	75.38 ± 1.34
VGG-Box-PCA	49.89 ± 1.97	84.21 ± 1.08
VGG-Box-KCSR	55.41 ± 1.41	87.00 ± 0.92
Navie Dynamic	86.00 ± 1.70	98.21 ± 1.08
Ours (Single-verif)	85.55 ± 1.30	96.31 ± 0.08
Ours (Dynamic MTL)	**87.30 ± 1.20**	**99.21 ± 1.07**

Best results in each column are in bold

Table 5 The evaluation of Photo to Caricature (P2C) identification (C2P) on dataset WebCaricature

Method	Rank-1 (%)	Rank-10 (%)
SIFT-Land-KCSR	23.42 ± 1.57	69.95 ± 2.34
VGG-Eye-PCA	36.18 ± 3.24	68.95 ± 3.25
VGG-Eye-KCSR	40.67 ± 3.61	75.77 ± 2.63
VGG-Box-PCA	50.59 ± 2.37	82.15 ± 1.31
VGG-Box-KCSR	55.53 ± 2.17	86.86 ± 1.42
Navie Dynamic	82.80 ± 1.60	97.81 ± 0.88
Ours (Single-verif)	81.70 ± 2.60	95.25 ± 1.08
Ours (Dynamic MTL)	**84.00 ± 1.60**	**99.01 ± 1.2**

Best results in each column are in bold

automatically according to the losses of tasks. However, the primitive dynamic method is stuck in the overtraining of the easy task due to the larger weight being always assigned to the easy task with smaller loss. This state can be hardly turned over unless the loss of the hard task can be occasionally decreased much quicker than the easy task and obtain a smaller loss than the one of the current easy task. This is why the primitive dynamic multi-task learning is stuck in the overtraining of the easy task and the undertraining of the hard task.

5.4 Multi-task learning for photo-caricature face recognition

In this section, we evaluate extensively the different approaches on the datasets CaVI and WebCaricature. Unlike the toy example, all tasks have been evaluated here. Figure 6 demonstrates the comparison of the proposed dynamic multi-task learning and the primitive dynamic multi-task learning for caricature identification, photo identification and the photo-caricature face verification on the two datasets. As

Fig. 7 The false positive (in the red rectangle) and false negative pairs obtained by our method for photo-caricature face verification from dataset CaVI

same as shown in the toy example, our proposed method can also adapt the weights of tasks enabling the network to focus on the training of the hardest task while the primitive dynamic method still preferentially train the easiest task. Besides, across all the datasets and the methods, the caricature identification task with a large loss is the hardest task to train. This is reasonable since the model has been pretrained on the images of photo, it is relative easy to train the photo related tasks rather than the caricature identification task which is trained nearly from scratch.

Table 2 shows the evaluation results for photo-caricature face verification, caricature identification and photo identification on dataset CaVI. It shows that the proposed dynamic multi-task learning method outperforms the state-of-art method CaVINet for all three tasks. We also evaluate the primitive dynamic multi-task learning method implemented based on our network. We can see that for the hard task, *i.e.,* caricature identification, the performance of the primitive dynamic multi-task learning (75.80%) is inferior to our method (85.61%) and also worse than the performance of

Fig. 8 The false positive (in the red rectangle) and false negative pairs obtained by our method for photo-caricature face verification from WebCaricature

the single-task model (78.20%), which proves that the primitive dynamic multi-task learning is incapable to well train the hard task.

In addition, we also report the performance of employing the caricature identification task with the photo identification model (V2C) and vice versa (C2V) as shown in Table 2. Comparing to single-task learning, it suggests that multi-task learning framework can obtain a much better performance by virtue of the sharing hidden layers which has learned the common features across the different recognition modalities. Comparing to C2V , V2C performs better since the photo identification model has been pretrained on the large dataset for photos, whereas the caricature model has been only trained on the relative small dataset for caricatures.

Tables 3, 4 and 5 demonstrate the evaluation results on the dataset WebCaricature. Since the methods proposed in [11] are the baseline methods for demonstrating the benchmark WebCaricature, the performance of our methods boost significant comparing to the baseline approaches. All the evaluations are conducted by the 10-folds cross validation

following the same evaluation protocol in [11]. We can see that our method achieves the best performance for all the evaluated tasks, *i.e.,* caricature-photo face verification task, caricature to photo identification (C2P) and photo to caricature (P2C) identification tasks. However, there is still much room to improve in terms of the validation rate (recall rate) at a low false accept rate (false positive rate).

5.5 Analysis

Figures 7 and 8 present some false positive and false negative pairs obtained by our method from the datasets CaVI and WebCaricature. The false positive pairs are the pairs with the different identities mistakenly recognized as the same person, while the false negative pairs are the pairs with same identities mistakenly recognized as the different persons. We can see that the caricatures and the photos in the false positive pairs (in the red rectangles) are similar to some extent, *e.g.,* the similar poses, facial expressions, hair styles, etc.. Nevertheless, the reason for resulting in the false negative pairs is diverse. The extreme distortion introduced by the exaggerated artistic style maybe the primary reason. What is interesting, human being can still perceive some delicate features to recognize the caricatures with exaggerated distortion, which indicates that maybe the machine can also learn to capture these features to improve the capacity for cross-modal photo-caricature face recognition.

6 Conclusion

In this work, we propose a dynamic multi-task learning approach for cross-modal photo-caricature face recognition. The proposed dynamic multi-task learning network can learn the feature representations of images from different modalities by combining different modality-specific tasks in the network. Rather than the existed multi-task learning methods, the proposed dynamic weight learning module can automatically generate the task weights enabling the network to focus on training the hard task instead of being stuck in the over-training of the easy task. Both the qualitative and quantitative analysis, and also the experimental results demonstrate the effectiveness of the proposed dynamic multi-task learning approach for cross-modal photo-caricature face recognition.

Although this dynamic multi-task learning approach is proposed for photo-caricature face recognition problems, the proposed architecture can be also easily extended to the other multi-task learning problem in the different domain, *e.g.,* employing handwritten recognition and word segmentation tasks with the proposed dynamic multi-task learning architecture.

References

1. Taigman, Yaniv, Yang, Ming, et al.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR, pp. 1701–1708, (2014)
2. Parkhi, Omkar M., Vedaldi, Andrea, Zisserman, Andrew, et al.: Deep face recognition. In: BMVC, p. 6, (2015)
3. Schroff, Florian, Kalenichenko, Dmitry, Philbin, James: Facenet: A unified embedding for face recognition and clustering. In: CVPR, pp. 815–823, (2015)
4. Liu, Weiyang, Wen, Yandong, Yu, Zhiding, Li, Ming, Raj, Bhiksha, Song, Le.: Sphereface: Deep hypersphere embedding for face recognition. In: The CVPR, vol. 1, p. 1 (2017)
5. Huang, Gary B., Ramesh, Manu, Berg, Tamara, Learned-Miller, Erik: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst (2007)
6. Wolf, Lior, Hassner, Tal, Maoz, Itay: Face recognition in unconstrained videos with matched background similarity. In: CVPR, 2011 IEEE Conference on, pp. 529–534. IEEE (2011)
7. Ahonen, Timo: Hadid, Abdenour, Pietikainen, Matti: face description with local binary patterns: application to face recognition. IEEE Transact. Pattern Anal. Mach. Intell. **28**(12), 2037–2041 (2006)
8. Tan, Xiaoyang, Triggs, Bill: Fusing gabor and lbp feature sets for kernel-based face recognition. In: International workshop on analysis and modeling of faces and gestures, pp. 235–249. Springer (2007)
9. Déniz, Oscar: Bueno, Gloria, Salido, Jesús, De la Torre, Fernando: Face recognition using histograms of oriented gradients. Pattern Recognit. Lett. **32**(12), 1598–1603 (2011)
10. Bicego, Manuele, Lagorio, Andrea, Grosso, Enrico, Tistarelli, Massimo: On the use of sift features for face authentication. In: Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on, pp. 35–35. IEEE (2006)
11. Huo, Jing, Li, Wenbin, Shi, Yinghuan, Gao, Yang, Yin, Hujun: Webcaricature: a benchmark for caricature recognition. In: British Machine Vision Conference (2018)
12. Mittal, Paritosh, Vatsa, Mayank, Singh, Richa: Composite sketch recognition via deep network-a transfer learning approach. In: 2015 International Conference on Biometrics (ICB), pp. 251–256. IEEE (2015)
13. Galea, Christian, Farrugia, Reuben A.: Forensic face photo-sketch recognition using a deep learning-based architecture. IEEE Signal Process. Lett. **24**(11), 1586–1590 (2017)
14. Li, Shan, Deng, Weihong: Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing (2020)
15. He, Ran, Wu, Xiang, Sun, Zhenan, Tan, Tieniu: Learning invariant deep representation for nir-vis face recognition. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
16. He, Ran, Xiang, Wu, Sun, Zhenan, Tan, Tieniu: Wasserstein cnn: learning invariant features for nir-vis face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1761–1773 (2018)
17. Kim, Donghyun, Hernandez, Matthias, Choi, Jongmoo, Medioni, Gérard: Deep 3d face identification. In: 2017 IEEE international joint conference on biometrics (IJCB), pp. 133–142. IEEE (2017)
18. Zulqarnain Gilani, Syed, Mian, Ajmal: Learning from millions of 3d scans for large-scale 3d face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1896–1905 (2018)
19. Garg, Jatin, Peri, Skand Vishwanath, Tolani, Himanshu, Krishnan, Narayanan C.: Deep cross modal learning for caricature verification and identification (cavinet). arXiv preprint arXiv:1807.11688, (2018)

20. Cai, Deng, He, Xiaofei, Han, Jiawei: Speed up kernel discriminant analysis. VLDB J. **20**(1), 21–33 (2011)
21. van der Maaten, Laurens, Hinton, Geoffrey: Visualizing data using t-sne. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
22. Ruder, Sebastian: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098, (2017)
23. Girshick, Ross: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448 (2015)
24. Ranjan, Rajeev, Patel, Vishal M., Chellappa, Rama: Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 121 (2017)
25. Tian, Yonglong, Luo, Ping, Wang, Xiaogang, Tang, Xiaoou: Pedestrian detection aided by deep learning semantic tasks. In: Proceedings of the CVPR, pp. 5079–5087 (2015)
26. Chen, Zhao, Badrinarayanan, Vijay, Lee, Chen-Yu, Rabinovich, Andrew: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. arXiv preprint arXiv:1711.02257, (2017)
27. Kendall, Alex, Gal, Yarin, Cipolla, Roberto: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7482–7491 (2018)
28. Yin, Xi, Liu, Xiaoming: Multi-task convolutional neural network for pose-invariant face recognition. IEEE Trans. Image Proces. **27**(2), 964–975 (2008)
29. Duong, Long, Cohn, Trevor, Bird, Steven, Cook, Paul: Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 845–850 (2015)
30. Misra, Ishan, Shrivastava, Abhinav, Gupta, Abhinav, Hebert, Martial: Cross-stitch networks for multi-task learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3994–4003 (2016)
31. Bragman, Felix J.S., Tanno, Ryutaro, Ourselin, Sebastien, Alexander, Daniel C., Cardoso, Jorge: Stochastic filter groups for multi-task cnns: Learning specialist and generalist convolution kernels. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1385–1394 (2019)
32. Chen, Weihua, Chen, Xiaotang, Zhang, Jianguo, Huang, Kaiqi: (2017) A multi-task deep network for person re-identification. In: AAAI, pp. 3988–3994
33. Zhang, Zhanpeng: Luo, Ping, Loy, Chen Change, Tang, Xiaoou, : Learning deep representation for face alignment with auxiliary attributes. IEEE Trans. Pattern Anal. Mach. Intell. **38**(5), 918–930 (2016)
34. Tran, Anh T., Nguyen, Cuong V., Hassner, Tal: Transferability and hardness of supervised classification tasks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1395–1405 (2019)
35. Sun, Yi, Wang, Xiaogang, Tang, Xiaoou: Deeply learned face representations are sparse, selective, and robust. In: CVPR, pp. 2892–2900 (2015)
36. Simonyan, Karen, Zisserman, Andrew: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. Wen, Yandong, Zhang, Kaipeng, Li, Zhifeng, Qiao, Yu: A discriminative feature learning approach for deep face recognition. In: European Conference on Computer Vision, pp. 499–515. Springer (2016)
38. Kemelmacher-Shlizerman, Ira, Seitz, Steven M., Miller, Daniel, Brossard, Evan: The megaface benchmark: 1 million faces for recognition at scale. In: Proceedings of the CVPR, pp. 4873–4882 (2016)
39. Zhang, Liliang, Lin, Liang, Wu, Xian, Ding, Shengyong, Zhang, Lei: End-to-end photo-sketch generation via fully convolutional representation learning. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 627–634 (2015)
40. Zhu, Jun-Yan, Park, Taesung, Isola, Phillip, Efros, Alexei A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232 (2017)
41. Wang, Lidan, Sindagi, Vishwanath, Patel, Vishal: High-quality facial photo-sketch synthesis using multi-adversarial networks. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pp. 83–90. IEEE (2018)
42. Saxena, Shreyas, Verbeek, Jakob: Heterogeneous face recognition with cnns. In: European conference on computer vision, pp. 483–491. Springer (2016)
43. Liu, Xiaoxiang, Song, Lingxiao, Wu, Xiang, Tan, Tieniu: Transferring deep representation for nir-vis heterogeneous face recognition. In: 2016 International Conference on Biometrics (ICB), pp. 1–8. IEEE (2016)
44. Lezama, José, Qiu, Qiang, Sapiro, Guillermo: Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6628–6637 (2017)
45. Collobert, Ronan, Weston, Jason: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning, pp. 160–167. ACM, (2008)
46. Deng, Li, Hinton, Geoffrey, Kingsbury, Brian: New types of deep neural network learning for speech recognition and related applications: An overview. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 8599–8603. IEEE (2013)
47. Szegedy, Christian, Ioffe, Sergey, Vanhoucke, Vincent, Alemi, Alexander A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
48. Simonyan, Karen, Omkar, M., et al. Parkhi. Fisher vector faces in the wild. In: BMVC, p. 4 (2013)
49. Amos, Brandon, Ludwiczuk, Bartosz, Satyanarayanan, Mahadev, et al. Openface: A general-purpose face recognition library with mobile applications. CMU School of Computer Science, 6, (2016)
50. MegviiInc. Face++ research toolkit. www.faceplusplus.com,. (December 2013)
51. Guo, Yandong, Zhang, Lei, Hu, Yuxiao, He, Xiaodong, Gao, Jianfeng: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision, pp. 87–102. Springer (2016)
52. Zhang, Kaipeng, Zhang, Zhanpeng, Li, Zhifeng, Qiao, Yu.: Joint face detection and alignment using multitask cascaded convolutional networks. Signal Proces. Lett. **23**(10), 1499–1503 (2016)
53. Glorot, Xavier, Bengio, Yoshua: Understanding the difficulty of training deep feedforward neural networks. In: 13th International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)