Learning from the Past: Meta-Continual Learning with Knowledge Embedding for Jointly Sketch, Cartoon, and Caricature Face Recognition

Wenbo Zheng
School of Software Engineering,
Xi'an Jiaotong University,
Xi'an, China.
State Key Laboratory for Management and Control of Complex Systems,
Institute of Automation,
Chinese Academy of Sciences,
Beijing, China.
zwb2017@stu.xjtu.edu.cn

Lan Yan
State Key Laboratory for Management and Control of Complex Systems,
Institute of Automation,
Chinese Academy of Sciences,
Beijing, China.
yanlan2017@ia.ac.cn

Fei-Yue Wang
State Key Laboratory for Management and Control of Complex Systems,
Institute of Automation,
Chinese Academy of Sciences,
Beijing, China.
feiyue.wang@ia.ac.cn

Chao Gou*
School of Intelligent Systems Engineering,
Sun Yat-sen University,
Guangzhou, China.
gouchao@mail.sysu.edu.cn

ABSTRACT

This paper deals with a challenging task of learning from different modalities by tackling the difficulty problem of jointly face recognition between abstract-like sketches, cartoons, caricatures and real-life photographs. Due to the significant variations in the abstract faces, building vision models for recognizing data from these modalities is an extremely challenging. We propose a novel framework termed as **Meta-Continual Learning with Knowledge Embedding** to address the task of jointly sketch, cartoon, and caricature face recognition. In particular, we firstly present a deep relational network to capture and memorize the relation among different samples. Secondly, we present the construction of our knowledge graph that relates image with the label as the guidance of our meta-learner. We then design a knowledge embedding mechanism to incorporate the knowledge representation into our network. Thirdly, to mitigate catastrophic forgetting, we use a meta-continual model that updates our ensemble model and improves its prediction accuracy. With this meta-continual model, our network can learn from its past. The final classification is derived from our network by learning to compare the features of samples. Experimental results demonstrate that our approach achieves significantly higher performance compared with other state-of-the-art approaches.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; **Machine learning approaches**; • **Applied computing** → *Media arts.*

KEYWORDS

Meta Learning; Heterogeneous Face Recognition; Sketch; Cartoon; Caricature; Continual Learning; Knowledge Embedding

ACM Reference Format:
Wenbo Zheng, Lan Yan, Fei-Yue Wang, and Chao Gou. 2020. Learning from the Past: Meta-Continual Learning with Knowledge Embedding for Jointly Sketch, Cartoon, and Caricature Face Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3394171.3413892

1 INTRODUCTION

Deep learning has been quite effective in narrowing the representational gap for multi-modal learning applications such as multi-modal or cross-modal face recognition. Prior work in this area focused on real-world facial images such as near-infrared, depth imagery, etc. Approaches for these modalities have been quite successful due to the inherent similarity in the structure of a face captured using different modalities [34, 42, 59]. However, multimedia facial analytics, where one of the modality is a sketch, cartoon, or caricature, is a challenging task due to the extreme levels of facial

*Corresponding Author

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413892

appearance variations (e.g., exaggerations, point of view, appearance, and the underlying artistic style) [27].

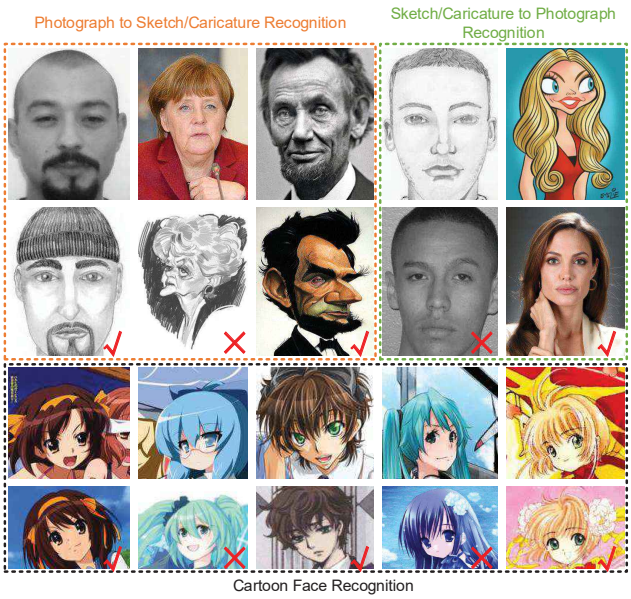

Figure 1: Illustration of The Heterogeneous Face Task. It consists of five task, i.e., photograph to sketch recognition, sketch to photograph recognition, photograph to caricature recognition, caricature to photograph recognition, cartoon face recognition. In the settings of this paper, we consider the above five tasks as a whole, i.e., the recognition of jointly sketch, cartoon, and caricature face.

Given the heterogeneous nature of photographs and its abstract faces (i.e., sketches/cartoons/caricatures) stemming from different generation mechanisms (i.e., intensity by digital sensor vs. drawing by hand) [10, 38, 56, 56], there can be large geometric deformations and texture differences between a face photograph and its associated abstract faces [19, 41, 57]. These factors make abstract face recognition a challenging heterogeneous face recognition problem [8]. In this paper, we focus on the recognition of jointly sketch, cartoon, and caricature face, as shown in Figure 1. In general, there are three main challenges for this heterogeneous face recognition:

(1) Many variations can influence recognition, such as facial appearance exaggerations and distortions, point of view, appearance, and the underlying artistic style.

(2) While joint training a model for sketch, cartoon, and caricature face recognition, the learning of the later tasks may degrade the performance of the models learned for the earlier tasks.

(3) Photographs and abstract faces in the datasets are limited. What's more, for these images, it seems to be implicitly related [29].

In contrast, even though there is a lot of variations in abstract faces beyond realism, humans are very good at recognizing the subjects. Why can human beings recognize abstract faces quickly and accurately with very little direct supervision or none at all? Probably because human beings can use the experience from the past to learn [44, 53–55, 58, 60], and the network can't. And isn't this one of the mechanisms of meta-learning [43]? We may use this mechanism to solve the above first issue. So, *why don't we use the principle of meta-learning to build a network for heterogeneous face recognition?*

Besides, our human brains seem to have this remarkable ability to learn lots of different tasks without any of them negatively interfering with each other [13, 33]. Continual learning algorithms try to achieve this ability for the neural networks and to solve the catastrophic forgetting problem [36]. We may use this algorithm to solve the above second issue. Thus, *why does not use continual learning for jointly training in heterogeneous face recognition?*

Furthermore, to address the third issue, facing limited information, humans still can learn to understand this scenario, due they acquire knowledge by integrating implicit relations [6]. In particular, they learn referents in knowledge by statistically matching words with occurrences of images in the environment [40]. Traditional recognition models, however, are usually developed based on single modalities or tasks with limited access to this implicit relations. Since a crucial aspect of traditional recognition model is to learn appropriate representations for designated tasks, it seems particularly important to combine implicit relations also in learning these representations. Isn't this exactly the problem that knowledge embedding [46] devote to solving? *Why not design knowledge embedding to embed the implicit relations among data into our model?*

To address the issues mentioned above, in this paper, we propose a novel meta-continual learning-based model for jointly sketch, cartoon, and caricature face recognition, which exploits knowledge embedding strategy in the whole process. We build a two-branch relation network via meta-continual learning. First, we use the embedding approach to do feature extraction of training images. In this process, we design knowledge embedding to guide our network. Then, to compare the features, we design a relation model that determines if they are from matching categories or not. Finally, to mitigate catastrophic forgetting, we design a meta-continual model that updates our whole model and improves the accuracy of its predictions. Experimental results show that our model performs better than similar works, and has strong robustness. The qualitative discussion suggests that the meta-learning-based proposed strategy achieves significantly higher performance compared with other meta-learning-based methods.

In summary, our main contributions are as follows:

☼ We propose a novel and unified approach to achieve jointly sketch, cartoon, and caricature face recognition. *To the best of our knowledge, this is the first attempt to study a jointly heterogeneous face recognition with respect to sketch, cartoon, and caricature simultaneously.* Experimental results show that the proposed approach has strong robustness and outperforms existing similar methods.

☼ We design a novel knowledge embedding mechanism to unify a knowledge graph with deep networks to facilitate heterogeneous faces recognition, with injecting the mined implicit relations among data into deep networks. *To the best our knowledge, this is the first attempt to study heterogeneous faces recognition method based on this knowledge embedding.*

☼ We present a novel meta-learning-based approach to learn the discriminative features cross different datasets.

☼ We design a novel strategy combing meta-learning with continual learning to learn how to guide the optimization of neural

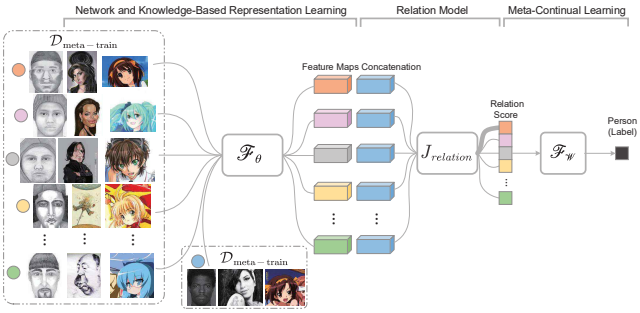

Figure 2: The Framework of Our Relation Network. It contains three modules: a network and knowledge-based representation learning model, a relation model, a meta-continual learning model. The network and knowledge-based representation learning model $\mathscr{F}_\theta$ parametrized by θ produces feature maps to represent feature extraction function. The relation model $J_{relation}(\cdot)$ represents the similarity between sample and query, which are from training set during the training phase, and from support set and query set, during the test phase, respectively. The meta-continual learning model $\mathscr{F}_\theta$ updated by weight $\mathscr{W}$ updates the whole model towards learning on new data quickly while minimizing forgetting, and produces the final results.

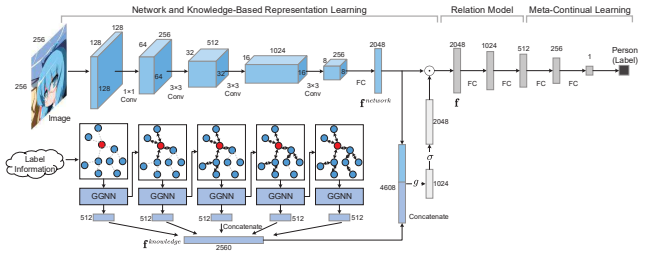

Figure 3: The Architecture of Our Relation Network. First of all, we use our network to extract the features of image, and get the image feature $\mathrm{f}^{network}$. Meanwhile, we construct our knowledge graph that relates images with corresponding label using label information to get knowledge embedding vector $\mathrm{f}^{knowledge}$. Then, we use the gated mechanism to fuse the $\mathrm{f}^{network}$ and $\mathrm{f}^{knowledge}$, and get the knowledge-based representation feature f. Finally, we apply the relation model and the meta-continual learning model.

network parameters for the problem of catastrophic forgetting, commonality and robustness cross different domains.

2 PROPOSED APPROACH

2.1 Problem Definition

We consider the problem of heterogeneous face recognition (sketch, cartoon, and caricature face recognition) as meta-continual classifier learning. The model is shown in Figure 2.

Notations The whole model consists of two phases: meta-training and meta-testing. In meta-training, our training data $\mathcal{D}_{\text{meta-train}} = \{(x_i, y_i)\}^n{}_{i=1}$ from a set of classes C_{train} are used for training a classifier, where x_i is a image, $y_i \in C_{train}$ is the corresponding person (label), and n is the number of training samples. In meta-testing, a support set of v labeled examples $\mathcal{D}_{\text{support}} = \{(x_j, y_j)\}^v{}_{j=1}$ from a set of new classes C_{test} is given, where x_j is a image for testing, and $y_j \in C_{test}$ is the corresponding person (label). The goal is to predict the labels of a query set $\mathcal{D}_{\text{query}} = \{(x_j)\}^{v+q}{}_{j=v+1}$, where q is the number of queries. This split strategy of training and support set aims to simulate the support and query set that will be encountered at test time. Further, we use the meta-learning on the training set to transfer the extracted knowledge to the support set. It aims to perform the model's learning on the support set better and classify the query set more successfully.

Learning to Continually Learn In the setting of meta-continual classifier learning, we divide the sake of traning representation learning into two goals, one is how to better obtain effective features and classification results for each task, the other is how to build on top of existing knowledge to learn on new data quickly while minimizing forgetting for all tasks. For the training set $\mathcal{D}_{\text{meta-train}}$, we define the sample set as $X = \{(x_i)\}^n{}_{i=1}$, where $X \in \mathcal{X} \subseteq \mathbb{R}^+$.

Our objective is to predict their regression labels, i.e., $Y = \{y_i\}^n{}_{i=1}$, where $Y \in \mathcal{Y}$. We achieve our goal by learning a mapping function $\mathscr{F} : \mathcal{X} \to \mathcal{Y}$. For each task, we define a mapping function $\mathscr{F}_\theta : \mathcal{X} \to \mathcal{Y}$ parametrized by θ to learn the discriminative features and achieve the classification effectively. If we regard all tasks as a whole, for all tasks, we define a mapping function $\mathscr{F}_\mathscr{W} : \mathcal{X} \to \mathcal{Y}$ updated by weight $\mathscr{W}$ to learn on new data quickly while minimizing forgetting. Obviously, these two functions composes the function $\mathscr{F}$: $\mathscr{F} = \mathscr{F}_\mathscr{W}(\mathscr{F}_\theta(X))$.

More concretely, given task $\mathscr{T}$ sampled from task distribution $p(\mathscr{T})$, let $Loss \in \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be the function that defines loss between a prediction y_j and target y_i as $Loss(y_j, y_i)$. If we assume that inputs X are seen proportionally to some density $\mu : \mathcal{X} \to [0, \infty)$, then we want to minimize the total loss $\mathscr{L}_{total}$ for all tasks:

$$\mathscr{L}_{total}(\mathscr{W}, \theta) = \mathbb{E}[Loss(\mathscr{F}(X), Y)] = \int [\int [Loss(\mathscr{F}(x), y)p(y|x)dy]]\mu(x)d(x) \quad (1)$$

where $\mathscr{W}$ and θ represent the set of parameters that are updated to minimize the total loss. To this end , we limit ourselves to learning by online updates on a single k length trajectory sampled from $p((\mathscr{S}_k|(\mathscr{T})$, where $\mathscr{S}_k = \{(x_i, y_i)\}^{i+k-1}{}_i(k < n + 1)$.

2.2 Knowledge Graph Construction and Representation

In this subsection, we construct our knowledge graph that relates images with corresponding person (label). For the construction of our knowledge graph, we use the GGNN [24] method. Built upon it, we use the GGNN to propagate node message through the graph and compute a feature vector for each node. All the feature vectors are then concatenated to generate the representation for the knowledge graph.

Principle of GGNN GGNN [24] is an end-to-end trainable network architecture that can learn features for arbitrary graph-structured data by iteratively updating node representation in a recurrent fashion. Formally, the input is a graph represented as

$\mathcal{G} = \{\mathbf{V}, \mathbf{A}\}$, in which $\mathbf{V}$ is the node set and $\mathbf{A}$ is the adjacency matrix denoting the connections among these nodes. We define t is the time step of conducting the knowledge graph. At $t = 0$, input feature vectors $\mathbf{x_v}$ that depends on the special task is initialized as the hidden state. Then, at time-step t, we define $\mathbf{h_v}^t$ as the hidden state. For each node $v \in \mathbf{V}$, the basic propagation recurrent process is formulated as

$$\begin{aligned} \mathbf{h}_v{}^0 &= \mathbf{x}_v \\ \mathbf{a}_v{}^t &= \mathbf{A}_v{}^T[\mathbf{h}_1{}^{t-1} \cdots \mathbf{h}_{|\mathbf{V}|}{}^{t-1}]^T + \mathbf{b} \\ \mathbf{h}_v{}^t &= gate(\mathbf{a}_v{}^t, \mathbf{h}_v{}^{t-1}) \end{aligned} \tag{2}$$

where $\mathbf{A}_v$ is a sub-matrix of $\mathbf{A}$ represents the connections of node v with its neighbors, and $gate$ denotes gated update mechanism, which is defined as:

$$\begin{aligned} \mathbf{z}_v{}^t &= \sigma(\mathbf{W}^z\mathbf{a}_v{}^t + \mathbf{U}^z\mathbf{h}_v{}^{t-1}) \\ \mathbf{r}_v{}^t &= \sigma(\mathbf{W}^r\mathbf{a}_v{}^t + \mathbf{U}^r\mathbf{h}_v{}^{t-1}) \\ \tilde{\mathbf{h}}_v^t &= \tanh(\mathbf{W}\mathbf{a}_v{}^t + \mathbf{U}(\mathbf{r}_v{}^t \odot \mathbf{h}_v{}^{t-1})) \\ \mathbf{h}_v{}^t &= (1 - \mathbf{z}_v{}^t) \odot \mathbf{h}_v{}^{t-1} + \mathbf{z}_v{}^t \odot \tilde{\mathbf{h}}_v^t \end{aligned} \tag{3}$$

where $\odot$, σ and tanh are the element-wise multiplication operation, the logistic sigmoid and hyperbolic tangent functions, respectively.

The propagation process is repeated until our fixed iteration T. During this process, we update the representation of each node based on its history state and the message sent by its neighbors. Thus, we can obtain the final hidden states $\{\mathbf{h}_1{}^T, \mathbf{h}_2{}^T, \ldots, \mathbf{h}_{|\mathbf{V}|}{}^T\}$. All in all, the computation process of equation (2) can be reduced to $\mathbf{h}_v{}^t = \text{GGNN}(\mathbf{h}_1{}^T, \mathbf{h}_2{}^T, \ldots, \mathbf{h}_{|\mathbf{V}|}{}^T; \mathbf{A}_v)$. Similar to [15], we employ an output network that is implemented by a fully-connected layer o, to compute node-level feature, expressed by

$$\mathbf{o}_v = o([\mathbf{h}_v{}^T, \mathbf{x}_v]), v = 1, 2, 3, \cdots |\mathbf{V}| \tag{4}$$

A Case of Constructing Image-Person GGNN Distinctly, we need to construct two knowledge graphs of which one relates images with corresponding person (label). Given dataset that covers C_{train} classes and n images, the graph has a node set $\mathbf{V}$ with $C_{train} + n$ elements. Similar to [4], we define the $C_{train} \times n$ matrix $\mathbf{S}_{Person\&Image}$ that denotes the confidence that this class has the image and its value range is $[0, 1]$. Then, we can get the adjacency matrix $\mathbf{A}_{Person\&Image}$ expressed as

$$\mathbf{A}_{Person\&Image} = \begin{bmatrix} \mathbf{0}_{C_{train} \times C_{train}} & \mathbf{S}_{Person\&Image} \\ \mathbf{0}_{C_{train} \times n} & \mathbf{0}_{n \times n} \end{bmatrix} \tag{5}$$

where $\mathbf{0}_{\cdot}$ is a zero vector with dimension $\cdot$.

Finally, by this way, we can get the knowledge graph $\mathcal{G}_{Person\&Image} = \{\mathbf{V}_{Person\&Image}, \mathbf{A}_{Person\&Image}\}$.

Knowledge Graph Representation After building the knowledge graph, we employ the GGNN to propagate node message through the graph and compute a feature vector for each node. All the feature vectors are then concatenated to generate the final representation for the knowledge graph.

We count the probabilities of all possible relationships given images and person in dataset, which are denoted $\mathbf{S} = \{s_0, s_1, \cdots, s_L\}$. We initialize the node refers to the image i with s_i, and the node refers to each person with a zero vector. Thus, we can get the input feature for each node can be represented as

$$\mathbf{x}_v = \begin{cases} [s_i, \mathbf{0}_{n-1}] & \text{if node } v \text{ refers to one image } i \\ [\mathbf{0}_n] & \text{if node } v \text{ refers to one person} \end{cases} \tag{6}$$

where $\mathbf{0}_{\cdot}$ is a zero vector with dimension $\cdot$.

After T iteration, according to the principle of the GGNN, we can get the node-level feature $\mathbf{o}_v{}^{Person\&Image}$ computed by Eq.(4). Finally, these features are concatenated to produce the final knowledge representation $\mathbf{f}^{knowledge}$.

2.3 Network-Based Representation Learning

Our network architecture is shown in Figure 3. Figure 3 describes a traditional process of convolution and pooling. We use the 6-layer network architecture. Taking an image as input, the output of the 6-th pooling layer is a 2048-dimensional vector, which we regard as network features. The kernels of network change in turns: $3 \times 256 \times 256 \rightarrow 128 \times 128 \times 128$ (Convolution, kernel size: 1×1) $\rightarrow 256 \times 64 \times 64$ (Convolution, kernel size: 3×3) $\rightarrow 512 \times 32 \times 32$ (Convolution, kernel size: 3×3) $\rightarrow 1024 \times 16 \times 16$ (Convolution, kernel size: 3×3) $\rightarrow 256 \times 8 \times 8$. Then, we apply the fully connected layer to change into 2048-dimensional vector, denoted as $\mathbf{f}^{network}$.

2.4 Knowledge-Based Representation Learning

We introduce the gated mechanism that embeds the knowledge representation to enhance the representation learning, considering suppressing non-informative features and allowing informational features to pass under the guidance of the our knowledge graph, similar to [5], we introduce a gated mechanism expressed as

$$\mathbf{f} = \sigma(g(\mathbf{f}^{network}, \mathbf{f}^{knowledge})) \odot \mathbf{f}^{network} \tag{7}$$

where σ is the logistic sigmoid, $\odot$ denotes the element-wise multiplication operation, g is a neural network that takes the concatenation of the feature of the final knowledge embedding and the feature of extracting by using the feature fusion network. It is implemented by two stacked fully connected layers.

2.5 Meta-learning Model

As illustrated in Figure 2, we define the function $\mathscr{F}_\theta$ represents feature extraction function using network and knowledge-based representation learning, i.e., the output of $\mathscr{F}_\theta$ is $\mathbf{f}$, and the function C represents feature concatenation function.

Relation Model Suppose sample $x_j \in \mathcal{D}_{\text{support}}$ and sample $x_i \in \mathcal{D}_{\text{meta-train}}$, the concatenated feature map of the training and testing sets is used as the relation model $J_{relation}(\cdot)$ to get a scalar in range of 0 to 1 representing the similarity between x_i and x_j, which is called relation score. Suppose we have one labeled sample for each of n unique classes, our model can generate n relation scores $Judge_{i,j}$ for the relation between one support input x_j and training sample set examples x_i:

$$Judge_{i,j} = J_{relation}(C(\overbrace{\mathscr{F}_\theta(x_i)}^{\text{The } \mathbf{f} \text{ of } x_i}, \overbrace{\mathscr{F}_\theta(x_j)}^{\text{The } \mathbf{f} \text{ of } x_j})) \quad i = 1, 2, \cdots, n \tag{8}$$

Furthermore, we do the operation of the element-wise sum over representation learning outputs of all samples from each training

class to form this class's feature map. This pooled class-level feature map is concatenated with the feature map of the test image as above.

Objective Function of Each Task We use mean square error (MSE) loss to train our model, regressing the relation score $Judge_{i,j}$ to the ground truth: matched pairs have similarity 1 and the mismatched pair have similarity 0.

$$Loss(\mathscr{F}_\theta(x_i; x_j), \{y_i; y_j\}) = \arg\min \sum_{i=1}^{n} \sum_{j=1}^{m} (Judge_{i,j} - (y_i == y_j))^2 \tag{9}$$

We design the two fully-connected layers to relation model. We use two fully-connected layers to have 1024 and 512 outputs, respectively, followed by a sigmoid function to get the final similarity scores mentioned in Eq. (9).

2.6 Meta-Continual Learning Model

After the process of learning the function $\mathscr{F}_\theta$, where θ is learned by minimizing *Loss* and then later fixed at meta-test time, we learn $\mathscr{F}_\mathscr{W}$ for $\mathscr{L}_{total}$ from a single trajectory $\mathscr{S}$ using fully online SGD updates in a single pass. We design the two fully-connected layers to meta-continual learning model treated as $\mathscr{W}$. We use two fully-connected layers to have 256 and 1 outputs, respectively.

Objective Function of All Tasks Therefore, our total objective is defined as:

$$\min_{\mathscr{W},\theta} \sum_{\mathscr{T}_i \sim p(\mathscr{T})} \mathscr{L}_{total}(\mathscr{W}, \theta) = \sum_{\mathscr{T}_i \sim p(\mathscr{T})} \sum_{\mathscr{S}_k{}^z \sim p(\mathscr{S}_k | \mathscr{T}_i)} [\mathscr{L}_{total_i}(U(\mathscr{W}, \theta, \mathscr{S}_k{}^z))] \tag{10}$$

where $\mathscr{S}_k{}^z = \{(x_i{}^z, y_i{}^z)\}^{i+k-1}{}_i$ and z is the statue (meta-train or support) of $\mathscr{S}_k$. $U(\mathscr{W}, \theta, \mathscr{S}_k{}^z)) = (\mathscr{W}_k, \theta)$ represents an update function where $\mathscr{W}_k$ is the weight vector after k steps of stochastic gradient descent.

3 EXPERIMENTAL RESULTS

In this section, we conduct experiments on three kind datasets (sketch datasets, cartoon datasets, and caricature datasets) to evaluate the performance of the proposed approach. Compared with the state-of-the-art models on these datasets, our approach yields better performance in term of accuracy (Acc.).

3.1 Experimental Setup

In this subsection, we describe the implementation details.

We resize the images from the three kind datasets to $256 \times 256 \times 3$. In sketch/caricature face recognition, we randomly choose 80 real face images and 80 corresponding sketches/caricatures of the 80 subjects for training, and the remaining samples of these 80 subjects are used in test. In cartoon face recognition, we randomly choose 16000 (8000×2) cartoon-face images of the 8000 subjects to construct a training set, and the remaining samples of these 8000 subjects are used as a test set. We randomly choose 10 times as per the above strategy and take the average recognition accuracy for comparison.

Knowledge Graph Construction and Representation For the GGNN model, the dimension of the hidden state is set as 4098 and that of the output feature is set as 512. The iteration time T is set as 5. GGNN is trained with ADAM following [28]. We get the 2560 (512×5) -dimensional knowledge embedding vector $\mathbf{f}^{knowledge}$.

Knowledge-Based Representation Learning We build two stacked fully connected layers in which the first one is 4608 (2048 + 2560) dimension to 1024 dimension followed by the hyperbolic tangent function while the second one is 1024 dimension to 2048.

Our Network Setting For all components, we use Adam optimizer [21] with a learning rate of 0.001 and a decay for every 50 epochs. We train 1000 epochs when the loss starts to converge.

3.2 Comparison with the State-of-the-Art Methods

We compare the state-of-the-art approaches with ours on these three kind datasets. In this subsection, "**Ours w/o GGNN**" means a variant of Ours, which only using network-based representation learning and not using knowledge graph and knowledge-based representation learning.

Sketch Face Recognition

We evaluate our approach on two tasks in sketch datasets: photograph to sketch recognition and sketch to photograph recognition on CUFS dataset [30, 47], CUFSF dataset [47, 52], IIIT-D Sketch dataset [2], PRIP-VSGC dataset [3, 18], PRIP-HDC dataset [23], MGDB dataset [35], UoM-SGFS dataset [16], and VIPSL dataset [37] respectively. The photograph to sketch recognition here is: given real faces of a public figure, we can recognize all the sketch faces of that public figure from a dataset of the sketch. Sketch to photograph recognition here is given sketch faces of that public figure, we can recognize all real face of a public figure from a dataset of the sketch.

Baselines for Sketch Face Recognition We compare against various state-of-the-art baselines for sketch face recognition, including CAL-HFR [25], DVR [50], DLFace [39], LightCNN+DVG [14], IACycleGAN [12], ASPT [51], and RCN [11].

Photograph to Sketch Recognition Table 1 shows the results of baselines and ours for photograph to sketch recognition.

Effect of Proposed Knowledge Embedding. For evaluating the impact of our approach, we compare results reported in row-"Ours w/o GGNN" and row-"Ours". Our method utilizes the same loss functions and features used in row-"Ours w/o GGNN" for a fair comparison. We observe that the proposed approach improves performance consistently in all cases. *It is evident that using knowledge embedding can enhance the effectiveness of our approach.*

Effect of Our Approach. From Table 1, it is evident that our approach is better than others. Specifically, ours is 11.6%, 15.1%, 19.8%, 24.8%, 33.3%, 28.1%, and 35.5% higher than CAL-HFR, DVR, DLFace, LightCNN+DVG, IACycleGAN, ASPT, and RCN, on the CUFS dataset, respectively. Besides, in other sketch datasets, there are similar scenarios as the above. *From above, our approach is more effective and robust than the state-of-the-arts approaches for the task of the photograph to sketch recognition.*

Sketch to Photograph Recognition Table 2 shows the results of baselines and ours for sketch to photograph recognition.

Effect of Proposed Knowledge Embedding. "Ours" is 9.4% higher than "Ours w/o GGNN" on the CUFS dataset. Besides, in other sketch datasets, there are similar scenarios as the CUFS dataset. *It shows the mechanism of knowledge embedding can improve the performance for the sketch to photograph recognition.*

Table 1: Recognition Accuracies (Percent) for The Task of Photograph to Sketch Recognition

Acc.(%) Method \ Dataset	CUFS	CUFSF	IIIT-D Sketch	PRIP-VSGC	PRIP-HDC	MGDB	UoM-SGFS	VIPSL
CAL-HFR	86.6%	82.4%	92.8%	78.0%	88.3%	92.0%	82.4%	67.2%
DVR	83.1%	79.9%	92.0%	76.3%	88.1%	91.0%	94.8%	56.3%
DLFace	78.5%	77.7%	90.8%	76.4%	82.8%	86.1%	94.8%	40.7%
LightCNN+DVG	73.5%	73.7%	97.0%	74.7%	87.0%	74.7%	93.0%	53.6%
IACycleGAN	65.0%	57.6%	88.4%	71.4%	83.7%	88.3%	89.4%	45.9%
ASPT	70.2%	59.0%	85.4%	66.9%	78.4%	67.0%	81.6%	45.4%
RCN	62.8%	72.8%	90.3%	65.6%	77.7%	65.5%	78.5%	37.6%
Ours w/o GGNN	*90.1%*	*89.8%*	*92.7%*	*88.7%*	*89.5%*	*90.0%*	*92.5%*	*72.4%*
Ours	**98.2%**	**91.2%**	**99.4%**	**92.6%**	**91.5%**	**93.7%**	**97.9%**	**74.6%**

Effect of Our Approach. "Ours" is better than others. Concretely, "Ours" is 10.3%, 11.8%, 18.8%, 26.1%, 30.0%, 25.3%, and 34.1% higher than CAL-HFR, DVR, DLFace, LightCNN+DVG, IACycleGAN, ASPT, and RCN, on the CUFS dataset, respectively. Besides, in other sketch datasets, there are similar scenarios as the above. *From above, our approach is more effective and robust than the state-of-the-arts approaches for the task of the sketch to photograph recognition.*

Table 2: Recognition Accuracies (Percent) for The Task of Sketch to Photograph Recognition

Acc.(%) Method \ Dataset	CUFS	CUFSF	IIIT-D Sketch	PRIP-VSGC	PRIP-HDC	MGDB	UoM-SGFS	VIPSL
CAL-HFR	89.5%	84.0%	95.0%	80.3%	89.5%	96.9%	84.6%	69.4%
DVR	88.0%	82.4%	94.0%	76.8%	90.3%	94.7%	97.8%	58.8%
DLFace	81.0%	82.2%	93.6%	77.6%	85.1%	87.9%	96.1%	43.0%
LightCNN+DVG	73.7%	73.8%	98.1%	78.7%	87.6%	78.4%	97.9%	57.1%
IACycleGAN	69.8%	61.4%	92.0%	76.1%	84.1%	92.6%	92.3%	48.5%
ASPT	74.5%	60.2%	89.3%	71.9%	82.0%	68.1%	86.1%	47.2%
RCN	65.7%	77.1%	91.8%	65.8%	78.0%	68.5%	82.0%	37.9%
Ours w/o GGNN	*90.4%*	*94.2%*	*92.9%*	*92.8%*	*90.6%*	*90.4%*	*97.2%*	*77.2%*
Ours	**99.8%**	**94.9%**	**100.0%**	**96.5%**	**91.5%**	**94.2%**	**99.9%**	**78.3%**

Caricature Face Recognition

We evaluate our approach on two tasks in caricature datasets: photograph to caricature recognition and caricature to photograph recognition on WebCaricature dataset [20], IIIT-CFW dataset [32], Caricature-207 dataset [22] and CaVI dataset [17] respectively. Photograph to caricature recognition here is: given real faces of a public figure, we can recognize all the caricature faces of that public figure from a dataset of caricature. Caricature to photograph recognition here is: given caricature faces of that public figure, we can recognize all real face of a public figure from a dataset of caricature.

Baselines for Caricature Face Recognition We compare against various state-of-the-art baselines for caricature face recognition, including PFRN [61], GFDF [7] and DDML [31].

Photograph to Caricature Recognition Table 3 shows the results of baselines and ours for photograph to caricature recognition.

Effect of Proposed Knowledge Embedding. From Table 3, "Ours" is 3.4% higher than "Ours w/o GGNN" on the WebCaricature dataset. Besides, in other caricature datasets, there are similar scenarios as WebCaricature. *It shows the design of knowledge embedding can improve the performance for the photograph to caricature recognition.*

Effect of Our Approach. "Ours" is better than others. At length, "Ours" is 6.2%, 8.1%, and 10.4% higher than GFDF, DDML, and PFRN, on the WebCaricature dataset, respectively. Besides, in other caricature datasets, there are similar scenarios as the above. *From above, our approach is more effective and robust than the state-of-the-arts approaches for the task of the photograph to caricature recognition.*

Table 3: Recognition Accuracies (Percent) for The Task of Photograph to Caricature Recognition

Acc.(%) Method \ Dataset	WebCaricature	IIIT-CFW	Caricature-207	CaVI
GFDF	87.5%	86.6%	87.5%	95.6%
DDML	85.7%	86.5%	86.7%	94.9%
PFRN	83.3%	84.5%	86.3%	96.3%
Ours w/o GGNN	*90.3%*	*92.4%*	*94.8%*	*96.8%*
Ours	**93.7%**	**97.6%**	**97.3%**	**99.8%**

Caricature to Photograph Recognition Table 4 shows the results of baselines and ours for caricature to photograph recognition.

Effect of Proposed Knowledge Embedding. From Table 4, "Ours" is 2.5% higher than "Ours w/o GGNN" on the WebCaricature dataset. Besides, in other caricature datasets, there are similar scenarios as WebCaricature. *Knowledge embedding plays an important role in our model for caricature to photograph recognition.*

Effect of Our Approach. "Ours" is better than others. Explicitly, "Ours" is 5.6%, 6.9%, and 4.3% higher than GFDF, DDML, and PFRN, on the WebCaricature dataset, respectively. Besides, in other caricature datasets, there are similar scenarios as the above. *From above, our approach is more effective and robust than the state-of-the-arts approaches for the task of caricature to photograph recognition.*

Table 4: Recognition Accuracies (Percent) for The Task of Caricature to Photograph Recognition

Acc.(%) Method \ Dataset	WebCaricature	IIIT-CFW	Caricature-207	CaVI
GFDF	91.1%	88.4%	91.7%	97.7%
DDML	89.9%	91.2%	90.5%	95.1%
PFRN	92.5%	93.7%	94.4%	95.7%
Ours w/o GGNN	*94.2%*	*93.9%*	*95.0%*	*98.5%*
Ours	**96.8%**	**99.9%**	**98.3%**	**99.9%**

Cartoon Face Recognition

We evaluate our approach on cartoon face recognition in the cartoon dataset (i.e., DanbooruCharacter dataset [1, 48]): given cartoon faces of a public figure, we can recognize other cartoon faces of that public figure from a dataset of the cartoon.

Baselines for Face Recognition We compare against various state-of-the-art baselines for face recognition, including Center-Loss Face [49], SphereFace [26], CosFace [45] and ArcFace [9].

Effect of Proposed Knowledge Embedding. From Table 5, "Ours" is 13.2% higher than "Ours w/o GGNN" on the DanbooruCharacter dataset. *It implies knowledge embedding is an important designing in our model for cartoon face recognition.*

Effect of Our Approach. From Table 5, "Ours" is better than others. More specially, "Ours" is 43.3%, 42.2%, 36.6%, and 30.5% higher than Center-Loss Face, SphereFace, CosFace, and ArcFace, on the DanbooruCharacter dataset, respectively. *From above, our approach is more effective and robust than the state-of-the-arts approaches for the task of cartoon face recognition.*

Table 5: Comparison for Cartoon Face Recognition

Acc.(%) Method \ Dataset	DanbooruCharacter
Center-Loss Face	24.5%
SphereFace	25.6%
CosFace	31.2%
ArcFace	37.3%
Ours w/o GGNN	*54.6%*
Ours	**67.8%**

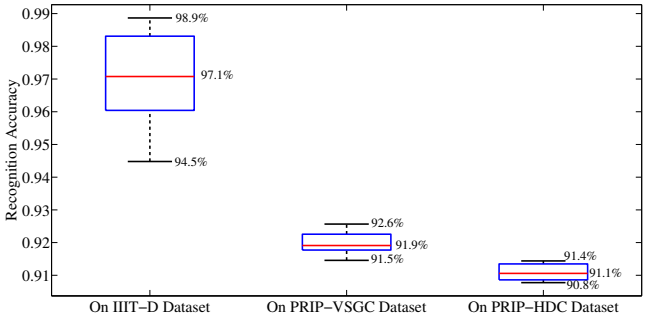

Figure 4: Cross-Database Recognition Accuracy of the Proposed Method. Our method is on the CUFS and CUFSF datasets for training, with respect to testing on the IIIT-D dataset, the PRIP-VSGC dataset and the PRIP-HDC dataset.

3.3 Discussion on the Generalization Ability

The data distributions in different heterogeneous face scenarios could be different from that during model development. To explore the generalization ability of the proposed method, we only use the CUFS and CUFSF datasets as sketch datasets to train our model for task of photograph to sketch recognition. Then, we evaluate it with cross-database testing on the IIIT-D dataset, the PRIP-VSGC composite sketch database, and the PRIP-HDC dataset. We run ten times following the above strategy in this discussion. In results, the recognition accuracy comparisons of testing on these datasets are shown in Figure 4. *This experiment indicates that the proposed method could achieve good recognition performance in such a challenging scenario.*

4 CONCLUSION

This paper presents a joint cross-modal model based on knowledge embedded meta-continual learning that can handle extreme variations present in sketches, cartoons, caricatures for recognition tasks. In particular, we present a novel deep relation network supervised via the knowledge embedding mechanism. To mitigate catastrophic forgetting, we design a meta-continual model that updates our network and improves the accuracy of its predictions. By this meta-continual model, our network can learn from its past. Our model that bridges sketches, cartoons, caricatures, and true-life face photograph modality facilitates the successful transfer of information across the modalities. Experimental results show our model has strong robustness and high recognition accuracy.

ACKNOWLEDGMENT

This work is supported in part by the Key Research and Development Program of Guangzhou (202007050002), in part by the National Natural Science Foundation of China (61806198, 61533019, U1811463), and in part by the National Key Research and Development Program of China (No. 2018AAA0101502).

REFERENCES

[1] Gwern Branwen Aaron Gokaslan Anonymous, the Danbooru community. 2019. Danbooru2018: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset. https://www.gwern.net/Danbooru2018. https://www.gwern.net/Danbooru2018 Accessed: DATE.

[2] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. 2012. Memetically Optimized MCWLD for Matching Sketches With Digital Face Images. *IEEE Transactions on Information Forensics and Security* 7, 5 (Oct 2012), 1522–1535. https://doi.org/10.1109/TIFS.2012.2204252

[3] IQ Biometrix. 2003. FACES 4.0. *Houston, TX: Author* (2003).

[4] Tianshui Chen, Liang Lin, Riquan Chen, Yang Wu, and Xiaonan Luo. 2018. Knowledge-Embedded Representation Learning for Fine-Grained Image Recognition. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 627–634. https://doi.org/10.24963/ijcai.2018/87

[5] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. 2019. Knowledge-Embedded Routing Network for Scene Graph Generation. In *Conference on Computer Vision and Pattern Recognition*.

[6] Michael W. Cole, Jeremy R. Reynolds, Jonathan D. Power, Grega Repovs, Alan Anticevic, and Todd S. Braver. 2013. Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience* 16, 9 (2013), 1348–1355. https://doi.org/10.1038/nn.3470

[7] Lingna Dai, Fei Gao, Rongsheng Li, Jiachen Yu, Xiaoyuan Shen, Huilin Xiong, and Weilun Wu. 2019. Gated Fusion of Discriminant Features for Caricature Recognition. In *Intelligence Science and Big Data Engineering. Visual Data Engineering*, Zhen Cui, Jinshan Pan, Shanshan Zhang, Liang Xiao, and Jian Yang (Eds.). Springer International Publishing, Cham, 563–573.

[8] T. de Freitas Pereira, A. Anjos, and S. Marcel. 2019. Heterogeneous Face Recognition Using Domain Specific Units. *IEEE Transactions on Information Forensics and Security* 14, 7 (July 2019), 1803–1816. https://doi.org/10.1109/TIFS.2018.2885284

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] Z. Deng, X. Peng, Z. Li, and Y. Qiao. 2019. Mutual Component Convolutional Neural Networks for Heterogeneous Face Recognition. *IEEE Transactions on Image Processing* 28, 6 (June 2019), 3102–3114. https://doi.org/10.1109/TIP.2019.2894272

[11] Zhongying Deng, Xiaojiang Peng, and Yu Qiao. 2019. Residual compensation networks for heterogeneous face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8239–8246.

[12] Yuke Fang, Weihong Deng, Junping Du, and Jiani Hu. 2020. Identity-aware CycleGAN for face photo-sketch synthesis and recognition. *Pattern Recognition* 102 (2020), 107249. https://doi.org/10.1016/j.patcog.2020.107249

[13] Martha J. Farah. 2018. Socioeconomic status and the brain: prospects for neuroscience-informed policy. *Nature Reviews Neuroscience* 19, 7 (2018), 428–438. https://doi.org/10.1038/s41583-018-0023-2

[14] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. 2019. Dual Variational Generation for Low-Shot Heterogeneous Face Recognition. In *NeurIPS*.

[15] Jianlong Fu, Heliang Zheng, and Tao Mei. 2017. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] C. Galea and R. A. Farrugia. 2016. A Large-Scale Software-Generated Face Composite Sketch Database. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 1–5. https://doi.org/10.1109/BIOSIG.2016.7736902

[17] Jatin Garg, Skand Vishwanath Peri, Himanshu Tolani, and Narayanan.C Krishna. 2018. Deep Cross Modal Learning for Caricature Verification and Identification (CaVINet). In *Proceedings of the 2018 ACM Conference on Multimedia*. ACM.

[18] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain. 2013. Matching Composite Sketches to Face Photos: A Component-Based Approach. *IEEE Transactions on Information Forensics and Security* 8, 1 (Jan 2013), 191–204. https://doi.org/10.1109/TIFS.2012.2228856

[19] S. Hu, N. Short, B. S. Riggan, M. Chasse, and M. S. Sarfraz. 2017. Heterogeneous Face Recognition: Recent Advances in Infrared-to-Visible Matching. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*. 883–890. https://doi.org/10.1109/FG.2017.126

[20] Jing Huo, Wenbin Li, Yinghuan Shi, Yang Gao, and Hujun Yin. 2018. WebCaricature: a benchmark for caricature recognition. In *British Machine Vision Conference*.

[21] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[22] B. F. Klare, S. S. Bucak, A. K. Jain, and T. Akgul. 2012. Towards automated caricature recognition. In *2012 5th IAPR International Conference on Biometrics (ICB)*. 139–146. https://doi.org/10.1109/ICB.2012.6199771

[23] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain. 2014. The FaceSketchID System: Matching Facial Composites to Mugshots. *IEEE Transactions on Information Forensics and Security* 9, 12 (Dec 2014), 2248–2263. https://doi.org/10.1109/TIFS.2014.2360825

[24] Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. 2016. Gated Graph Sequence Neural Networks. In *Proceedings of ICLR'16* (proceedings of iclr'16 ed.). https://www.microsoft.com/en-us/research/publication/gated-graph-sequence-neural-networks/

[25] D. Liu, X. Gao, N. Wang, J. Li, and C. Peng. 2020. Coupled Attribute Learning for Heterogeneous Face Recognition. *IEEE Transactions on Neural Networks and Learning Systems* (2020), 1–14. https://doi.org/10.1109/TNNLS.2019.2957285

[26] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[27] J. Lu, V. E. Liong, and J. Zhou. 2018. Simultaneous Local Binary Feature Learning and Encoding for Homogeneous and Heterogeneous Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 8 (Aug 2018), 1979–1993. https://doi.org/10.1109/TPAMI.2017.2737538

[28] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2017. The More You Know: Using Knowledge Graphs for Image Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[29] Julian McAuley and Jure Leskovec. 2012. Image Labeling on a Network: Using Social-Network Metadata for Image Classification. In *Computer Vision – ECCV 2012*. 828–841.

[30] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. 1999. XM2VTSDB: The extended M2VTS database. In *Second international conference on audio and video-based biometric person authentication*, Vol. 964. 965–966.

[31] Z. Ming, J. Burie, and M. Muzzamil Luqman. 2019. Dynamic Deep Multi-task Learning for Caricature-Visual Face Recognition. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 1. 92–97. https://doi.org/10.1109/ICDARW.2019.00021

[32] Ashutosh Mishra, Shyam Nandan Rai, Anand Mishra, and C. V. Jawahar. 2016. IIIT-CFW: A Benchmark Database of Cartoon Faces in the Wild. In *Computer Vision – ECCV 2016 Workshops*, Gang Hua and Hervé Jégou (Eds.). Springer International Publishing, Cham, 35–47.

[33] Yohsuke R. Miyamoto, Shengxin Wang, and Maurice A. Smith. 2020. Implicit adaptation compensates for erratic explicit strategy in human motor learning. *Nature Neuroscience* (2020). https://doi.org/10.1038/s41593-020-0600-3

[34] Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, Xueming Li, Chen Change Loy, and Xiaogang Wang. 2016. A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution. *Image and Vision Computing* 56 (2016), 28 – 48. https://doi.org/10.1016/j.imavis.2016.09.001

[35] Shuxin Ouyang, Timothy M. Hospedales, Yi-Zhe Song, and Xueming Li. 2016. ForgetMeNot: Memory-Aware Forensic Facial Sketch Matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[36] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks* 113 (2019), 54 – 71. https://doi.org/10.1016/j.neunet.2019.01.012

[37] Chunlei Peng, Xinbo Gao, Nannan Wang, and Jie Li. 2018. Face recognition from multiple stylistic sketches: Scenarios, datasets, and evaluation. *Pattern Recognition* 84 (2018), 262 – 272. https://doi.org/10.1016/j.patcog.2018.07.014

[38] Chunlei Peng, Xinbo Gao, Nannan Wang, and Jie Li. 2019. Sparse graphical representation based discriminant analysis for heterogeneous face recognition. *Signal Processing* 156 (2019), 46 – 61. https://doi.org/10.1016/j.sigpro.2018.10.015

[39] Chunlei Peng, Nannan Wang, Jie Li, and Xinbo Gao. 2019. DLFace: Deep local descriptor for cross-modality face recognition. *Pattern Recognition* 90 (2019), 161 – 171. https://doi.org/10.1016/j.patcog.2019.01.041

[40] Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsay, Kenneth D. Miller, Richard Naud, Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema, João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording. 2019. A deep learning framework for neuroscience. *Nature Neuroscience* 22, 11 (2019), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

[41] H. Roy and D. Bhattacharjee. 2016. Local-Gravity-Face (LG-face) for Illumination-Invariant and Heterogeneous Face Recognition. *IEEE Transactions on Information Forensics and Security* 11, 7 (July 2016), 1412–1424. https://doi.org/10.1109/TIFS.2016.2530043

[42] Shreyas Saxena and Jakob Verbeek. 2016. Heterogeneous Face Recognition with CNNs. In *Computer Vision – ECCV 2016 Workshops*, Gang Hua and Hervé Jégou (Eds.). Springer International Publishing, Cham, 483–491.

[43] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. 2018. Learning to Compare: Relation Network for Few-Shot Learning. In *Computer Vision and Pattern Recognition (CVPR)*.

[44] Flood Sung, Li Zhang, Tao Xiang, Timothy M. Hospedales, and Yongxin Yang. 2017. Learning to Learn: Meta-Critic Networks for Sample Efficient Learning. *CoRR* abs/1706.09529 (2017). arXiv:1706.09529 http://arxiv.org/abs/1706.09529

[45] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[46] Q. Wang, Z. Mao, B. Wang, and L. Guo. 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Transactions on Knowledge and Data Engineering* 29, 12 (Dec 2017), 2724–2743. https://doi.org/10.1109/TKDE.2017.2754499

[47] X. Wang and X. Tang. 2009. Face Photo-Sketch Synthesis and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 11 (Nov 2009), 1955–1967. https://doi.org/10.1109/TPAMI.2008.222

[48] Yan Wang. 2019. Danbooru 2018 Anime Character Recognition Dataset. https://github.com/grapeot/Danbooru2018AnimeCharacterRecognitionDataset. https://github.com/grapeot/Danbooru2018AnimeCharacterRecognitionDataset

[49] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A Discriminative Feature Learning Approach for Deep Face Recognition. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 499–515.

[50] Xiang Wu, Huaibo Huang, Vishal M Patel, Ran He, and Zhenan Sun. 2019. Disentangled variational representation for heterogeneous face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9005–9012.

[51] S. Yu, H. Han, S. Shan, A. Dantcheva, and X. Chen. 2019. Improving Face Sketch Recognition via Adversarial Sketch-Photo Transformation. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*. 1–8. https://doi.org/10.1109/FG.2019.8756563

[52] W. Zhang, X. Wang, and X. Tang. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR 2011*. 513–520. https://doi.org/10.1109/CVPR.2011.5995324

[53] Wenbo Zheng, Chao Gou, and Fei-Yue Wang. 2020. A novel approach inspired by optic nerve characteristics for few-shot occluded face recognition. *Neurocomputing* 376 (2020), 25 – 41. https://doi.org/10.1016/j.neucom.2019.09.045

[54] Wenbo Zheng, Chao Gou, and Lan Yan. 2019. A Relation Hashing Network Embedded with Prior Features for Skin Lesion Classification. In *Machine Learning in Medical Imaging*, Heung-Il Suk, Mingxia Liu, Pingkun Yan, and Chunfeng Lian (Eds.). Springer International Publishing, Cham, 115–123.

[55] Wenbo Zheng, Chao Gou, Lan Yan, and Shaocong Mo. 2020. Learning to Classify: A Flow-Based Relation Network for Encrypted Traffic Classification. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/3366423.3380090

[56] W. Zheng, C. Gou, L. Yan, and F. Wang. 2019. Differential-Evolution-Based Generative Adversarial Networks for Edge Detection. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2999–3008.

[57] W. Zheng, L. Yan, C. Gou, and F. Wang. 2020. Graph Attention Model Embedded With Multi-Modal Knowledge For Depression Detection. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6.

[58] Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. [n.d.]. Federated Meta-Learning for Fraudulent Credit Card Detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization.

[59] Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. [n.d.]. JND-GAN: Human-Vision-Systems Inspired Generative Adversarial Networks for Image-to-Image Translation. *GAN* 50 ([n. d.]), 1.

[60] Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. 2020. Webly Supervised Knowledge Embedding Model for Visual Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[61] W. Zheng, L. Yan, C. Gou, W. Zhang, and F. Wang. 2019. A Relation Network Embedded with Prior Features for Few-Shot Caricature Recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. 1510–1515. https://doi.org/10.1109/ICME.2019.00261