OPEN-SET RECOGNITION FOR FACIAL-EXPRESSION RECOGNITION

Mihiro Uchida, Shota Orihashi, Akihiko Takashima, Yoshihiro Yamazaki, Ryo Masumura

NTT Computer and Data Science Laboratories, NTT Corporation, Japan

ABSTRACT

We address distinguishing whether an input is a facial image by learning only a facial-expression recognition (FER) dataset. To avoid misclassification in FER, it is necessary to distinguish whether the input is a facial image. Unfortunately, collecting exhaustive non-face images is costly. Therefore, distinguishing whether the input is a facial image by learning only an FER dataset is important. A representative method for this task is learning reconstruction of only facial images and determining high-error samples between input images and reconstructed images as non-face images. However, reconstruction is difficult on facial images because such images contain detailed features. Our key idea to tackle the task without reconstruction is assuming that facial images will match several emotions, whereas non-face images will not match any emotion. Therefore, we propose a method for training a discriminator that determines whether the inputs and emotions match using counterfactual pairs in an FER dataset. A metric for the task is then obtained by taking into account each emotion in the posterior probability that inputs and emotions match, estimated by the discriminator. Experiments on the RAF-DB dataset vs. the Stanford Dogs dataset and AffectNet datasets showed the effectiveness of our method.

***Index Terms*—** Open-set recognition, Facial-expression recognition, Projection discriminator

1. INTRODUCTION

Facial-expression recognition (FER) is a task of classifying input human facial images into several classes of emotion. FER using facial expression classifiers on the basis of deep neural networks (DNNs) has been extensively studied [1–7] due to DNNs' high classification ability [8–10]. Unfortunately, DNNs cannot correctly predict classes for unlearned input. In other words, the facial-expression classifier misclassifies if a non-face image is input to it. To avoid misclassification, we should distinguish whether facial images or non-face images are input into the facial-expression classifier. The same can be said even when using face detection as preprocessing to crop the facial area from the whole image. This is because non-face images that are unlearned images are input to the facial-expression classifier if face detection failed, i.e., over-detection or false-detection. Thus, recognizing facial expressions and distinguishing facial images or non-face images at the same time is important. This task is called open-set recognition (OSR) [11] for FER. OSR is a task of simultaneously solving the classification of images of a learned class and distinguishing whether an input is an unlearned-class image. Assuming that the learned-class image is a facial-expression image, and the unlearned-class image is a non-face image, OSR can be applied to FER.

Previous studies on OSR investigated methods using outputs of a classification task [12–16]. These methods use classification to compute the probability of how precisely the input belongs to a learned-class. It is assumed that the probability of a class is high for learned-class images, whereas the probability of all classes is low for unlearned-class images. Unfortunately, the probability is low for both unlearned-class images and class-ambiguous images. The class-ambiguous images occur in facial expressions (e.g., expressions between fear and angry), as shown in Fig. 1, whereas images handled in ordinal classification tasks consists only class-precise images (e.g., dogs or cats). This makes it difficult to distinguish facial images or non-face images.

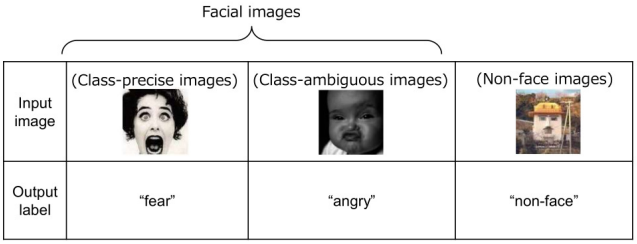

Fig. 1. Open-set recognition for facial expression recognition.

Other OSR studies investigated methods for learning image reconstruction [17–19]. These methods use an OSR metric, such as reconstruction error or similarity of features, which is extracted through learning reconstruction assuming that only learned-class images can be reconstructed. These methods can distinguish facial images and non-face images even if facial images include class-ambiguous images because the classes do not appear when applying the OSR metric. However, it is difficult to learn to reconstruct only facial images despite assuming that only learned-class images can be reconstructed because facial images input into the facial-expression classifier are complex. Even if facial images can be reconstructed, it is possible for all images, including non-face images, to be reconstructed. This means that it is difficult to distinguish facial images or non-face images with the OSR metrics. Thus, OSR for FER requires a method for handling even complex and class-ambiguous images.

To distinguish facial images including class-ambiguous images and non-face images, we develop OSR metrics, with which classes do not appear using a method but for reconstruction. We can apply these metrics if classes are eliminated in a class-conditioned probability that denotes whether an image is a facial image estimated using a discriminative model. We construct a discriminative model assuming that a non-face image will not match any emotion class, whereas a facial image including a class-ambiguous image is considered to match at least one emotion class. This can be useful to handle facial images and non-face image differently. If we estimate the class-conditioned probabilities of whether the input images match any emotion, we can apply these OSR metrics.

We propose a method of OSR for FER for eliminating classes from estimated class-conditioned probabilities to evaluate whether an input image and class match. To estimate the class-conditioned probabilities, we apply a projection discriminator [20] of a class-

conditional generative adversarial network (GAN) that models the class-conditioned probability of input samples by conditioning the class with an inner product operation to OSR for FER. This projection discriminator learns that the input is a real image or generated image from a given input image and class. We develop a modified projection discriminator to be trained on the assumption that the input image matches the ground-truth class, and the input image does not match the other classes as a binary classification task from given feature maps of input extracted from a pre-trained facial-expression classifier and given class.

We conducted experiments to evaluate the OSR for FER performance of the proposed method using Real-world Affective Faces Database (RAF-DB) [21], an FER dataset, and Stanford Dogs [22], an image-classification dataset of dog breeds. We used Stanford Dogs because dogs are considered likely to be captured in images at the same time as humans. We also evaluated the OSR for FER performance of the proposed method using AffectNet [23], which contains both facial-expression images and non-face images. We found through these experiments that the proposed method performs better on the area under the receiver operating characteristic (AUROC) curve than other methods that learn reconstructions or use classification outputs.

The contributions of this study are as follows. (1) We focus on OSR for FER that is necessary to handle complex and class-ambiguous images for the first time to the best of our knowledge, (2) present the proposed method that uses our developed projection discriminator to achieve OSR for FER by assuming that a non-face image will not match any emotion class, whereas a class-ambiguous image is considered to match at least one emotion class, and discuss our experiments to evaluate the OSR for FER performance of the proposed method.

2. OPEN-SET RECOGNITION FOR FACIAL-EXPRESSION RECOGNITION

Facial-expression recognition: Let $\boldsymbol{x}_i$ be the input image and $y_i = \{1, \cdots, K\}$ be the corresponding emotion-class label, then the training dataset can be represented as $\mathcal{D}_{\text{train}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N_{\text{train}}}$, where N_{train} is the number of samples in $\mathcal{D}_{\text{train}}$ and K is the number of emotion-class labels. Note that $\mathcal{D}_{\text{train}}$ consists of only facial images. The DNN-based facial-expression classifier is then constructed through training on $\mathcal{D}_{\text{train}}$. The classifier determines the estimated emotion-class label $\hat{y}_i$ as follows:

$$\hat{y}_i = \arg\max_y P(y|\boldsymbol{x}_i; \boldsymbol{\theta}_{\text{C}}), \tag{1}$$

$$P(y|\boldsymbol{x}_i; \boldsymbol{\theta}_{\text{C}}) = \boldsymbol{f}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\text{C}}), \tag{2}$$

where $\boldsymbol{\theta}_{\text{C}}$ are the trainable parameters of this DNN-based facial-expression classifier $\boldsymbol{f}(\cdot)$.

Open-set recognition for facial-expression recognition: Let the dataset of facial images be denoted as $\mathcal{D}_{\text{face}} = \{(\boldsymbol{x}_i)\}_{i=1}^{N}$, where N is the number of images in $\mathcal{D}_{\text{face}}$. Let the dataset of non-face images be denoted as $\mathcal{D}_{\text{non-face}} = \{(\boldsymbol{x}_i)\}_{i=1}^{M}$, where M is the number of images in $\mathcal{D}_{\text{non-face}}$. OSR involves inputting samples from a dataset that mixes images from facial images and non-face images $\mathcal{D}_{all} = \mathcal{D}_{\text{face}} \cup \mathcal{D}_{\text{non-face}}$ during inference. It also involves predicting the label $\bar{y}_i$ of $\boldsymbol{x}_i$ with a classifier $\boldsymbol{g}(\boldsymbol{x}_i)$ as follows:

$$\bar{y}_i = \boldsymbol{g}(\boldsymbol{x}_i) = \begin{cases} \hat{y}_i, & \text{if } \boldsymbol{x}_i \in \mathcal{D}_{\text{face}}, \\ K+1, & \text{if } \boldsymbol{x}_i \in \mathcal{D}_{\text{non-face}}, \end{cases} \tag{3}$$

Note that $K+1$ means that the input image is a non-face image.

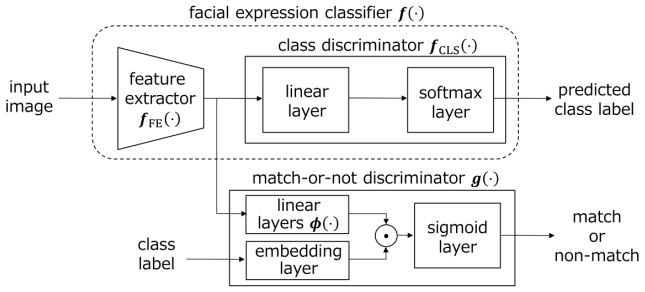

Fig. 2. Overview of proposed method. Note that inference can be done with any class label.

Whether $\boldsymbol{x}_i$ belongs to $\mathcal{D}_{\text{face}}$ or $\mathcal{D}_{\text{non-face}}$ is determined using one of two OSR metrics:$h_{\text{face}}(\cdot)$, which is related to $P(o = \text{face}|\boldsymbol{x}_i)$, or $h_{\text{non-face}}(\cdot)$, which is related to $P(o = \text{non-face}|\boldsymbol{x}_i)$, where $o \in \{\text{face}, \text{non-face}\}$. Whether $\boldsymbol{x}_i$ belongs to $\mathcal{D}_{\text{face}}$ is determined as follows:

$$\begin{cases} \boldsymbol{x}_i \in \mathcal{D}_{\text{face}}, & \text{if } h_{\text{face}}(\boldsymbol{x}_i) > \rho_{\text{face}}, \\ \boldsymbol{x}_i \in \mathcal{D}_{\text{non-face}}, & \text{otherwise}, \end{cases} \tag{4}$$

or

$$\begin{cases} \boldsymbol{x}_i \in \mathcal{D}_{\text{face}}, & \text{if } h_{\text{non-face}}(\boldsymbol{x}_i) < \rho_{\text{non-face}}, \\ \boldsymbol{x}_i \in \mathcal{D}_{\text{non-face}}, & \text{otherwise}, \end{cases} \tag{5}$$

where ρ_{face} and $\rho_{\text{non-face}}$ are the thresholds to determine if an image is a facial image or non-face image.

3. PROPOSED METHOD

An overview of the proposed method is shown in Fig. 2. The proposed method consists of a feature extractor, class discriminator, and match-or-not discriminator. The feature extractor computes feature maps to input into the class discriminator and match-or-not discriminator. The class discriminator computes $\hat{y}_i$ for Eq. (3). It computes h_{face} and $h_{\text{non-face}}$ to determine whether the input into the feature extractor is a facial image or non-face image. The training of the proposed method consists of the following two steps: training the feature extractor and the class discriminator as a facial-expression classifier to handle complex images, and training match-or-not discriminator to obtain an OSR metric to handle class-ambiguous images. To correctly recognize a class-ambiguous image as a face, we use the features extracted from the classifier to obtain metrics related to $P(o = \text{face}|\boldsymbol{x}_i)$ or $P(o = \text{non-face}|\boldsymbol{x}_i)$ by eliminating a class label y in the "probability that $\boldsymbol{x}_i$ and y match". This section is divided into three sections: one discussing the feature extractor and class discriminator, another discussing the match-or-not discriminator, and the other discussing the OSR metrics.

Training feature extractor and class discriminator: We train a model for FER to extract features and predict emotion classes. Using the feature extractor $\boldsymbol{f}_{\text{FE}}(\cdot)$ and class discriminator $\boldsymbol{f}_{\text{CLS}}(\cdot)$, the posterior probability for each class can be expressed as follows:

$$P(y|\boldsymbol{x}_i; \boldsymbol{\theta}_{\text{C}}) = \boldsymbol{f}_{\text{CLS}}(\boldsymbol{f}_{\text{FE}}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\text{FE}}); \boldsymbol{\theta}_{\text{CLS}}), \tag{6}$$

where $\boldsymbol{\theta}_{\text{C}} = \{\boldsymbol{\theta}_{\text{FE}}, \boldsymbol{\theta}_{\text{CLS}}\}$ are the trainable parameters of each model. Note that this equation can be obtained by substituting $\boldsymbol{f}(\cdot) = \boldsymbol{f}_{\text{CLS}}(\boldsymbol{f}_{\text{FE}}(\cdot))$ into Eq. 2, so we can obtain $\hat{y}$ by substituting Eq. (6) into Eq. (1).

The $\boldsymbol{\theta}_{\text{C}}$ is trained using the following loss function:

$$\mathcal{L}_{\text{C}} = -\frac{1}{N_{\text{train}}} \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_{\text{train}}} \log P(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}_{\text{C}}). \tag{7}$$

Learning match-or-not discriminator: We consider $P(m|\boldsymbol{x}_i, y)$ as a metric related to $P(o = \text{face}|\boldsymbol{x}_i)$ or $P(o = \text{non-face}|\boldsymbol{x}_i)$ to eliminate y, where $m \in \{\text{match}, \text{non-match}\}$. In other words, we consider a task that takes $\boldsymbol{x}_i$ and y as inputs and estimates "whether $\boldsymbol{x}_i$ and y are matched." This task can be expressed as follows:

$$P(m|\boldsymbol{x}_i, y; \boldsymbol{\theta}_{\mathrm{D}}) = \boldsymbol{g}(\boldsymbol{x}_i, y; \boldsymbol{\theta}_{\mathrm{D}}), \tag{8}$$

where $\boldsymbol{g}(\cdot)$ denotes the binary classifier to estimate whether $\boldsymbol{x}_i$ and y matches, and $\boldsymbol{\theta}_{\mathrm{D}}$ denotes the trainable parameters of $\boldsymbol{g}(\cdot)$.

The $\boldsymbol{\theta}_{\mathrm{D}}$ is trained using the following loss function.

$$\begin{aligned}\mathcal{L}_{\mathrm{D}} = &-\frac{1}{N_{\mathrm{train}}} \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_{\mathrm{train}}} \log P(m = \text{match}|\boldsymbol{x}, y; \boldsymbol{\theta}_{\mathrm{D}}) \\ &-\frac{1}{(K-1) \times N_{\mathrm{train}}} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}_{\mathrm{cf}}} \log P(m = \text{non-match}|\boldsymbol{x}_i, yi; \boldsymbol{\theta}_{\mathrm{D}}),\end{aligned} \tag{9}$$

where counterfactual dataset $\mathcal{D}_{\mathrm{cf}} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{(K-1)\times N_{\mathrm{train}}}$ is the set of an image and non-matched class label, and K is the number of emotion classes. We construct $\mathcal{D}_{\mathrm{cf}}$ by pairing $\boldsymbol{x}_i$ and all class labels but the ground-truth class label in $\mathcal{D}_{\mathrm{train}}$.

We use the projection discriminator [20] used in class-conditional GAN [24] as $\boldsymbol{g}(\cdot)$. This is because OSR for FER corresponds to the task in which $P(r|\boldsymbol{x}_i, y)$ learned in the class-conditional GAN is changed from $r \in \{\text{real}, \text{fake}\}$ to $m \in \{\text{match}, \text{non-match}\}$.

We describe the process of the projection discriminator, which takes the inner product of the feature extracted from the input and embedded emotion-class labels. This inner product is used to solve the binary classification problem. We use $\boldsymbol{f}_{\mathrm{FE}}(\cdot)$ as input into the projection discriminator to handle even complex images. This can be expressed as follows:

$$\boldsymbol{g}(\boldsymbol{x}_i, y; \boldsymbol{\theta}_{\mathrm{D}}) = \mathrm{Sigmoid}(\boldsymbol{\phi}(\boldsymbol{f}_{\mathrm{FE}}(\boldsymbol{x}_i); \boldsymbol{\theta}_{\phi}) \cdot \mathbf{Embed}(y; \boldsymbol{\theta}_{\mathrm{Embed}})), \tag{10}$$

where $\boldsymbol{\theta}_{\mathrm{D}} = \{\boldsymbol{\theta}_{\phi}, \boldsymbol{\theta}_{\mathrm{Embed}}\}$ are trainable parameters, $\cdot$ denotes the inner product operation, $\mathrm{Sigmoid}(\cdot)$ is the sigmoid function, $\boldsymbol{\phi}(\cdot)$ is a DNN consisting of linear layers, and $\mathbf{Embed}(\cdot)$ is the embedding function that embeds a class label into a vector. When updating $\boldsymbol{\theta}_D$, we keep $\boldsymbol{\theta}_C$ constant. Note that the inference using $\boldsymbol{g}(\boldsymbol{x}_i, y)$ can be executed even if y is not given in $\mathcal{D}_{\mathrm{all}}$. This is because any emotion class can be input into $\boldsymbol{g}(\boldsymbol{x}_i, y)$. The detail of inference are described in next.

Obtaining the metrics for open-set recognition: We can obtain the OSR metrics by eliminating y in the output of $\boldsymbol{g}(\boldsymbol{x}_i, y; \boldsymbol{\theta}_{\mathrm{D}})$. We use the following two methods to obtain these metrics: empirical and marginal. For both methods, we input all emotion classes into $\boldsymbol{g}(\boldsymbol{x}_i, y)$ to obtain the OSR metrics.

We first describe the marginal method. To eliminate y in the output of $\boldsymbol{g}(\boldsymbol{x}_i, y; \boldsymbol{\theta}_{\mathrm{D}})$, we can use marginalization as follows:

$$\begin{cases} h_{\text{face}}(\boldsymbol{x}_i) = \sum_y P(m = \text{match}|\boldsymbol{f}_{\mathrm{FE}}(\boldsymbol{x}_i), y) P(y), \\ h_{\text{non-face}}(\boldsymbol{x}_i) = \sum_y P(m = \text{non-match}|\boldsymbol{f}_{\mathrm{FE}}(\boldsymbol{x}_i), y) P(y). \end{cases} \tag{11}$$

We also describe the empiric method. Intuitively, a class is considered to be non-face images if the posterior probability of the class that matches the best is low. Therefore, we apply the following OSR metrics.

Table 1. Dataset structures.

Method	face or not	train	test
RAF-DB [21]	face	12,270	3,067
Stanford Dogs [22]	non-face	–	8,580
facial images in AffectNet [23]	face	310,969	4,500
non-face images in AffectNet [23]	non-face	–	500

$$\begin{cases} h_{\text{face}}(\boldsymbol{x}_i) = \max_y P(m = \text{match}|\boldsymbol{f}_{\mathrm{FE}}(\boldsymbol{x}_i), y), \\ h_{\text{non-face}}(\boldsymbol{x}_i) = 1 - \max_y P(m = \text{match}|\boldsymbol{f}_{\mathrm{FE}}(\boldsymbol{x}_i), y). \end{cases} \tag{12}$$

This can be considered the intuitively approximated method of the marginal methods.

4. EXPERIMENTS

4.1. Experiment on facial-expression recognition

In this experiment, we evaluated the OSR for FER performance of the proposed and comparison methods for handling complex and class-ambiguous images.

Datasets: We evaluated OSR for FER in the following two experimental settings: using RAF-DB [21] vs. Stanford Dogs [22], and facial images vs non-face images in AffectNet [23]. In RAF-DB vs. Stanford Dogs, we carried out OSR of images in the "happy", "angry", "sad", "fear", "surprised", "disgust", and "neutral" classes, and dog images in Stanford Dogs, which are regarded as non-face images. In the facial images vs. non-face images in AffectNet, we carried out OSR of images in the "happy", "angry", "sad", "fear", "surprised", "contempt", "disgust", "neutral", and "uncertain" classes, and images in the "non-face" class from AffectNet. The "non-face" class includes images such as of sculptures, paintings, and other non-face objects. In both settings, we used images of facial expressions as the learned class, and non-face images as the unlearned class when we evaluated previous studies. Table 1 shows the structure of the datasets.

Evaluation: the OSR performance of the proposed and comparison methods was evaluated using the AUROC curve. This is to evaluate the goodness of the OSR metric regardless of the threshold or balance between facial images and non-face images. Therefore, we did not set ρ_{face} or $\rho_{\text{non-face}}$ to evaluate each method.

The following five methods were used for comparison.

- CROSR [18]: The classification problem and image reconstruction are learned simultaneously. If the features extracted from the classification head and the features extracted from the image reconstruction do not belong to the distribution of any learned class image, the method determines them as unlearned.
- C2AE [17]: The reconstruction of the input image is learned from the intermediate outputs of the network and class labels. When the input image and class label match, the reconstruction error is learned so that the reconstruction error with the input image is small, and when they do not match, the reconstruction error with any image of the input class is learned so that the reconstruction error with any image of the input class is small. The method determines the input as unlearned when the reconstruction error becomes large for all classes.
- Softmax [13]: This method uses the output of a classification. The posterior probability of each class is assumed smaller

when the input is an image of an unlearned class. This method determines that the input with the smallest posterior probability is an image of an unlearned class.

- Mahalanobis distance [14]: This method uses the intermediate output of a classification. Images in the learned classes are assumed to follow a class-conditional Gaussian distribution in the space of pre-softmax. Inputs that have a low probability of belonging to any class-conditional Gaussian distribution are determined to be images of the unlearned class.
- OpenMax [12]: This method uses the intermediate output of a classification. The features of images in the unlearned class are assumed to not belong to the distribution formed by the features of the learned class. Images that do not belong to the distribution of the learned classes are determined to be images of the unlearned class.
- Ours: This method models $P(m|\boldsymbol{x}_i, y)$ using the intermediate outputs of a classification. The details are described in Section 3. We used the OSR metrics $h_{\text{non-face}}(\cdot)$ in Eq. (11) and Eq. (12) as the OSR metrics.

The experiments were conducted by fixing the classification models handled with each method to a well-known light DNN-based classifier, MobileNetv3 [25] and a well-known high accuracy DNN-based classifier, EfficientNet [10]. The architecture of the projection discriminator used with the proposed method was implemented as the inner product of the outputs of the three linear layers and the class labels embedded by the embedding layer, as in a previous study [20]. In previous studies, decoders of the image-reconstruction model passed the output of the layer with varying map sizes of the intermediate features of the classification model [17, 18]. The decoder implemented an up-sampling layer, transposed convolution layer, and batch-normalization layer [26], as in previous studies [17, 18], so that the passed features could be used to generate images of the same size as the input. The transposed convolution layer was activated using the rectified linear unit (ReLU) [27].

The facial-expression classifier used for each method was a common model trained on each FER dataset. Other models were trained until convergence. The Adam optimizer was used for all training. Adam's parameters followed those in a previous study [28].

Results: The experimental results are shown in Table 2. Ours achieved higher performance than CROSR and C2AE, which learn reconstruction. It also achieved higher performance than Softmax, Mahalanobis distance, and OpenMax, which use the output of classification. These results suggest that Ours can execute OSR for FER for handling complex and class-ambiguous images. The results indicate that there is little difference in performance betweeen the method to apply OSR metrics in Eq. (11) and Eq. (12).

4.2. Ablation study

In this study, we estimated $P(m|\boldsymbol{x}_i, y)$ using a projection discriminator as an OSR metric. In this experiment, we verified the following three class-conditioning method for estimating $P(m|\boldsymbol{x}_i, y)$.

- summation: The sum of $\boldsymbol{f}_{\text{FE}}(\boldsymbol{x}_i)$ and $\mathbf{Embed}(y)$ is input to the match-or-not discriminator for estimating $P(m|\boldsymbol{x}_i, y)$.
- concatenation: A vector combining $\boldsymbol{f}_{\text{FE}}(\boldsymbol{x}_i)$ and $\mathbf{Embed}(y)$ is input to the match-or-not discriminator for estimating $P(m|\boldsymbol{x}_i, y)$.
- Feature-wise Linear Modulation (FiLM) [29]: FiLM creates a weight matrix and bias matrix for each class and linearly transforms $\boldsymbol{f}_{\text{FE}}(\boldsymbol{x}_i)$ with the given class into feature maps. The feature map is input into the match-or-not discriminator for estimating $P(m|\boldsymbol{x}_i, y)$.

Table 2. OSR for FER performance. Best under each condition of setting is depicted in **bold.**

Model	Method	RAF-DB vs Stanford Dogs	AffectNet
Mobile -Netv3 [25]	C2AE [17]	0.504	0.501
	CROSR [18]	0.553	0.553
	Softmax [13]	0.806	0.516
	Mahalanobis distance [14]	0.635	0.574
	OpenMax [12]	0.833	0.540
	Ours with Eq. (11)	0.839	0.629
	Ours with Eq. (12)	**0.843**	**0.638**
Efficient -Net [10]	C2AE [17]	0.722	0.710
	CROSR [18]	0.513	0.512
	Softmax [13]	0.848	0.801
	Mahalanobis distance [14]	0.582	0.574
	OpenMax [12]	0.627	0.540
	Ours with Eq. (11)	0.816	0.798
	Ours with Eq. (12)	**0.854**	**0.802**

Table 3. Results of comparing class-conditioning methods. Best is depicted in **bold.**

Method	AUROC
summation	0.801
concatenate	0.804
FiLM [29]	0.816
projection discriminator	**0.843**

The architecture of each match-or-not discriminator model was a three-layer neural network, and the model for obtaining features was the same MobileNetv3. We compared each method for RAF-DB vs. Stanford Dogs. The OSR metric was $h_{\text{non-face}}$ obtained using Eq. (12).

The experimental results are listed in Table 3. The projection discriminator achieved higher performance than the other methods. These results indicate that the projection discriminator is suitable for modeling $P(m|\boldsymbol{x}_i, y)$.

5. CONCLUSION

We proposed a method using a match-or-not discriminator for OSR for FER. OSR for FER is a challenging task that handles complex images and class-ambiguous images. With the proposed method, it is assumed that a non-face image does not match any emotion class, whereas a facial image including the class-ambiguous image matches at least one emotion class. We trained the match-or-not discriminator to predict whether the emotion class matches the input image using the intermediate output of the facial-expression classifier. The output of this match-or-not discriminator was applied as an OSR metric by eliminating emotion classes. Experiments showed that the proposed method performs better than previous methods in OSR for FER for two dataset settings: RAF-DB vs. Stanford Dogs and AffectNet. We also experimentally verified that the projection discriminator is the best class-conditioning method for the proposed method.

6. REFERENCES

[1] Naima Otberdout, Anis Kacem, Mohamed Daoudi, Lahoucine Ballihi, and Stefano Berretti, "Deep covariance descriptors for facial expression recognition," *arXiv preprint arXiv:1805.03869*, 2018.

[2] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool, "Covariance pooling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 367–374.

[3] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, et al., "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 543–550.

[4] Takuhiro Kaneko, Kaoru Hiramatsu, and Kunio Kashino, "Adaptive visual feedback generation for facial expression improvement with multi-task deep neural networks," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 327–331.

[5] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image and Vision Computing*, vol. 65, pp. 66–75, 2017.

[6] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 443–449.

[7] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 118–126.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[10] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[11] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2013.

[12] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1563–1572.

[13] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *5th International Conference on Learning Representations*, 2017.

[14] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, pp. 7167–7177, 2018.

[15] Shu Kong and Deva Ramanan, "Opengan: Open-set recognition via open data generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 813–822.

[16] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman, "Open-set recognition: A good closed-set classifier is all you need," in *International Conference on Learning Representations*, 2021.

[17] Poojan Oza and Vishal M Patel, "C2ae: Class conditioned auto-encoder for open-set recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2307–2316.

[18] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura, "Classification-reconstruction learning for open-set recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4016–4025.

[19] He Zhang and Vishal M Patel, "Sparse representation-based open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1690–1696, 2016.

[20] Takeru Miyato and Masanori Koyama, "cgans with projection discriminator," in *International Conference on Learning Representations*, 2018.

[21] Shan Li, Weihong Deng, and JunPing Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[22] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*. Citeseer, 2011, vol. 2.

[23] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[24] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[25] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al., "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[26] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[27] Abien Fred Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.