Self-Supervised Learning for Facial Action Unit Recognition through Temporal Consistency

Liupei Lu
liupeilu@usc.edu
Leili Tavabi
ltavabi@usc.edu
Mohammad Soleymani
soleymani@ict.usc.edu

Institute for Creative Technologies
University of Southern California
Playa Vista, CA
USA

Abstract

Facial expressions have inherent temporal dependencies that can be leveraged in automatic facial expression analysis from videos. In this paper, we propose a self-supervised representation learning method for facial Action Unit (AU) recognition through learning temporal consistencies in videos. To this end, we use a triplet-based ranking approach that learns to rank the frames based on their temporal distance from an anchor frame. Instead of manually labeling informative triplets, we randomly select an anchor frame along with two additional frames with predefined distances from the anchor as positive and negative. To develop an effective metric learning approach, we introduce an aggregate ranking loss by taking the sum of multiple triplet losses to allow pairwise comparisons between adjacent frames. A Convolutional Neural Network (CNN) is used as encoder to learn representations by minimizing the objective loss. We demonstrate that our encoder learns meaningful representations for AU recognition with no labels. The encoder is evaluated for AU detection on various detasets including BP4D, EmotioNet and DISFA. Our results are comparable or superior to the state-of-the-art AU recognition through self-supervised learning.

1 Introduction

Facial expressions play an important role in social communication of emotions and intentions [28]. The Facial Action Coding System (FACS) is a taxonomy of facial activities that can describe expressions by anatomical action units, *e.g.*, chick raiser (AU 6) [6]. Unlike facial expressions, *e.g.*, happiness, sadness, FACS provides an objective measure for describing all facial expressions [20]. Supervised learning for recognition of facial action units requires laborious manual labeling by trained coders. To alleviate the problem of the scarcity and cost of labeling, semi-supervised and self-supervised learning methods have been proposed to leverage unlabeled data for facial action unit recognition [13, 18, 38]. Self-supervised representation learning leverages proxy supervision, which has great potential in improving performance in computer vision applications and video analysis [8, 13, 18, 22].

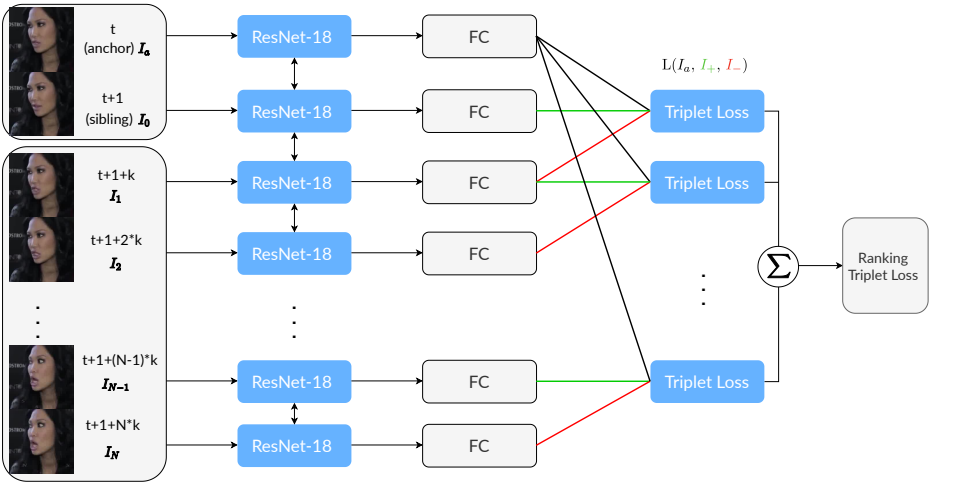

Figure 1: Proposed parallel encoders network takes a sequence of frames extracted from a video. The anchor frame is selected at time t, the sibling frame at $t+1$, and the following frames at equal intervals from $t+1+k$ to $t+1+Nk$. All input frames are fed to ResNet-18 [10] encoders with shared weights, followed by a fully-connected layer to generate 256d embeddings. L2-norm is applied on output embeddings. We then compute triplet losses for adjacent frame pairs along with the fixed anchor frame. In each adjacent pair, the preceding frame is the positive sample and the following frame is the negative sample. Finally, all triplet losses are added to form the ranking triplet loss.

Existing self-supervised methods for facial expression analysis leverage reconstruction loss between the generated and the target frames through an AutoEncoder to learn a representation that encodes facial activities [13, 18]. In this paper, we focus on learning intrinsic temporal dependencies for facial expressions. Facial activities, by nature, are temporally consistent, and this consistency can be used for representation learning.

In order to learn the existing temporal consistency, we use a temporally sensitive triplet-based metric learning, a self-supervised proxy loss, to learn to rank sequences of neighboring frames from videos in correct temporal order. Learning to rank through triplet loss involves training an encoder that learns to make the distance between the anchor and the positive sample smaller than the distance between the anchor and the negative sample. The majority of triplet-based metric learning methods use labels to indicate positive and negative classes. Instead of manually labeling informative triplets, we take advantage of the consistency of the facial behavior in consecutive frames from videos.

We introduce triplet ranking loss, a combination of a sequence of N triplet losses. The input to the losses has two parts: (1) an anchor frame and (2) a sequence sampled at constant time intervals with length of $N+1$. For each video, we first randomly select a frame as the anchor frame for all N triplets. The sequence is selected by taking the anchor's adjacent frame, which we refer to as the sibling frame, along with the sibling's following N frames sampled at a constant interval k. Every individual triplet pair compares the distance of two consecutive frames with reference to the anchor, where the earlier frame corresponds to the positive sample and the following frame as the negative. This method encodes the internal temporal consistency within the sequence. The sum of N triplet losses enables the network to learn to sort the frames in the correct order. We choose ResNet18 [10] as our encoder to generate compact embeddings from face images and train the encoder to minimize the sum

of triplet ranking losses. An overview of the proposed method is available in Fig. 1.

We evaluate our learned embeddings in downstream classifications tasks of AU detection and expression classification using multiple datasets in Section 4. The code and the trained model weights are shared for the sake of reproducibility[1]. The main contributions of this work are as follows. (i) We propose to use temporal consistency in facial activities for self-supervised representation learning. (ii) A temporal ranking triplet loss is designed that can distill the temporal consistency through pairwise comparisons. (iii) We demonstrate the effectiveness of this simple, yet powerful, model for facial action unit recognition.

2 Related Work

There is a well-established line of research for representation learning for facial expression analysis [20]. Neural networks have been used to learn rich local and global representations to capture facial attributes [1, 15, 26].

Sequential Learning. Sequential representation learning has been used for learning representation in a variety of domains including speech processing, robotic path planning and Natural Language Processing (NLP) [2, 8, 14, 22, 23]. Misra *et al*. [22] focuses on sequential data from video frames and leverages spatiotemporal signals to learn visual representations using unsupervised methods. They use sequential verification for learning representations by predicting the temporal validity of a given randomly shuffled sequence of frames. Chu *et al*. [2] looks at the facial AUs by jointly modeling the spatial and temporal representations. They extract spatial representations using a CNN and feed the representations to Long Short-Term Memory recurrent neural networks (LSTMs) to model temporal dynamics. However, the method proposed in [22] only focuses on learning general macro-motions from video instead of specific facial movements. Chu et al. [2] focuses specifically on facial AU detection but is trained in a supervised fashion, which relies on manual labeling and tends to overfit to a specific database [6]. The most related work in sequential learning uses Dynamic Representation (DR) [33], which infers dynamic representations that can summarize motion from static images. They propose a rank loss to capture a bi-directional flow through time from a sequence of frames centered at input image. However, DR calculates scores to rank all possible combinations of pairs given a sequence, whereas our ranking triplet loss captures inherent temporal consistency between consecutive frames. In addition, DR generates a single dynamic image representation through encoder-decoder architecture, and focuses on learning the symmetry in image domain. In contrast, our method adopts metric learning by capturing the temporal consistency in the embedding space which can produce compact embeddings for AU analysis.

Self-supervised Learning. There is a growing interest in self-supervised learning due to its ability to leverage existing structure in data to adopt supervisory signals for generating labels, and therefore decreasing the need for expensive manual labels [12, 13, 18, 27, 36]. Existing work on this topic including FAb-Net [13] and TCAE [18] use facial movements from videos as the supervisory signal by depicting such movements as the transformation between two face images in different frames. FAb-Net learns to map a source frame to a target frame by training an AutoEncoder with a reconstruction loss. Inspired by FAb-Net, TCAE models the movements by separating them into two representations, including smaller movements (facial actions) and larger movements (head poses) between the source and the

[1] https://git.io/JJSI6

target frames. They use a Twin Cycle Autoencoder to disentangle the movements related to head pose from action units. Since both self-supervised methods learn to extract embeddings through discrete pixel reconstruction, they do not focus on learning the temporal information conveyed in videos.

Deep Metric Learning. Metric learning has been widely studied in a variety of domains [9, 30, 32, 34, 35]. The goal of metric learning is to capture similarities between data points in the embedding space, usually achieved by applying a proxy loss directly on the learned embeddings. There are a variety of proxy losses proposed, such as contrastive loss [9], triplet loss [30, 34], N-pair-mc loss [32], and ranked-list loss [35]. Traditional contrastive loss and triplet loss consider one pair of positive and negative sample each time, and force the distance between the anchor and the positive sample to be smaller than the anchor and the negative sample. While most of triplet loss learning focuses on label-based triplets, FECNet [34] utilizes triplet loss to learn visual similarity between expressions. They manually annotate expression triplets where two expressions in each triplet are more similar to each other than the third one. FECNet succeeds in image retrieval and expression classification from the learned embeddings. Our encoder architecture is close to FECNet, however, contrary to their use of manual labels, we label triplets using frames' temporal order. In addition, similar to N-pair-mc loss, our proposed method is a structured loss which involves multiple negative samples instead of one. Although unlike N-pair-mc loss, our method has the capability of ranking the distance of each negative sample from the anchor.

3 Method

3.1 Problem Formulation

To learn the temporal consistency from videos, we define a loss function by adding multiple triplet losses, each measuring the pairwise distance between adjacent frames to the anchor. For each sequence of frames extracted from a video, let I_a be the fixed anchor frame randomly selected at time t, and the sibling frame I_0 be the anchor's adjacent frame at $t+1$. Starting from the sibling frame, we draw a sequence of frames $\{I_0, I_1, I_2, ...I_N\}$ with size $N+1$. Each frame I_n is extracted at the time $t+1+kn$, where k is a fixed interval between frames in the sequence. The encoder takes the anchor frame and the following sequence as input and encodes corresponding embeddings $\{x^a, x^0, x^1, ..., x^N\}$. We provide the detailed explanation on training through the combination of N triplet losses in Section 3.3.

3.2 Triplet Loss

Traditional triplet loss is a distance metric loss which measures the difference between the distances of positive x^+ and negative x^- samples from an anchor in the embedding space. The objective is to minimize the distance between the anchor and the positive sample, x^+, and maximize the distance between the anchor and the negative sample x^-. The triplet loss function is defined as: $L(I_a, I_+, I_-) = \max(0, d(x^+, x^a) - d(x^-, x^a) + \delta)$, where $d()$ denotes the distance function. The hinge function with margin δ ensures the loss will not reach zero unless the difference between the distances of the negative and positive sample from the anchor is greater than δ. In this paper, we use L2-norm to measure the Euclidean distance between embeddings. Although the triplet loss increases the distance of the negative samples from the anchor in each update, it is unclear to what extent this distance should be

increased. Consequently, the traditional triplet loss generally experiences slow convergence and requires mining of non-trivial triplet samples to accelerate training process [32].

3.3 Temporal Ranking Triplet Loss

To alleviate slow convergence of networks with a single triplet loss, we use the temporal order to rank distances between a sequence of negative samples and a unique anchor with multiple triplet losses. Therefore, we leverage the smooth and continuous changes of the facial expressions in short periods of time as our temporally consistent signal.

Inspired by N-pair-mc loss [32], we allow multiple negative samples inside one loss function as follows.

$$\mathcal{L}\left(\left\{x,x^{+},\{x_i\}_{i=1}^{N-1}\right\};f\right)=\log\left(1+\sum_{i=1}^{N-1}\exp\left(f^{\top}f_i-f^{\top}f^{+}\right)\right) \tag{1}$$

We compute the sum of individual triplet losses between adjacent frames in the sequence. Given a set of input embeddings $\{x^a,x^0,x^1,...,x^N\}$ and L2-distance, d, our temporal ranking triplet loss is defined as follows:

$$L\left(\left\{x^a,\{x^n\}_{n=0}^{N}\right\};d\right)=\sum_{n=1}^{N}\max\left(0,d(x^{n-1},x^a)-d(x^n,x^a)+\delta\right) \tag{2}$$

One significant difference between our method and the N-pair-mc is that our positive sample is not fixed. In the N-pair-mc loss, all negative samples are compared with the same positive sample. This approach is not adequate for our sampled frames, since distances between the frames and the anchor are gradually increasing. Therefore, we switch to pairwise comparisons between consecutive frames at fixed intervals. Every nth triplet involves consecutive frame pairs from the sequence, where x^{n-1} is the positive sample and the following frame x^n is the negative sample. We introduce the sibling frame x^0 as the positive sample at $n=1$, in order to achieve consistent formulation for the entire sequence. This ranking loss ensures each frame has a greater distance to the anchor in the embedding space compared to the distance of its preceding frame with respect to the anchor. In addition, instead of using multi-class logistic loss, we directly take the sum of hinge losses. Hinge loss results in the individual triplet loss becoming zero if the distance between the negative and positive samples are larger than δ.

3.4 Network Architecture

To accelerate the convergence of the ranking loss, we use a ResNet18 architecture [10] pretrained on ImageNet [4] as encoder. ResNet18 was selected due to its competitive results as shown by previous work on facial expression analysis with transfer learning using pretrained weights [19]. In addition, we have also evaluated the performance of a slimmer network MobileNet_V2 [29] and a denser network DenseNet121 [11], results are available in supplementary material. The results indicate that a simple and relatively shallow architecture, such as ResNet18 has sufficient capacity for this task, making it a balanced choice between performance and complexity.

We remove the last fully connected layer from the ResNet18, taking the output from the last max-pooling layer with 512 hidden units. We add a linear (fully connected) layer to generate a compact 256-dimensional embedding. An L2 normalization is applied to the output to normalize the embeddings. For a sequence of frame inputs $\{I_a,I_0,I_1,...,I_N\}$, as defined in

Section 3.1, we use encoders with shared weights as shown in Fig. 1. The stacked encoders encode frames to embeddings $\{x^a, x^0, x^1, ..., x^N\}$ simultaneously, and feed them into temporal ranking triplet loss, as described in Section 3.3. While training the encoders at a fixed interval k, different intervals could encode different facial features, with different temporal granularity. To take advantage of these facial features at different granularities, we propose an ensemble model by concatenating multiple embeddings generated from independently trained encoders at different intervals (k).

4 Experiments

4.1 Training Setup

Our network is trained on VoxCeleb2 video dataset [4], which consists of 150k videos from around 6,000 speakers. We use pre-cropped videos provided by VoxCeleb2 at 25 frames per second (fps). The frames are then resized and center-cropped in the same fashion as the setup in [18]. During training, jittering is used to augment the input data same as [18]. For each sampled video frame sequence, we draw an anchor frame, a sibling frame and a sequence of N frames as described in Section 3.1. These video sequences are randomly split for train and test by a 80/20 ratio in video-independent fashion. The model is trained in PyTorch framework, with a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.001 and a momentum of 0.9. We use the margin size $\delta = 0.03$ in equation 2, and $N = 10$ for the number of triplets in the sequence. Furthermore, we evaluate our method using different choices of intervals including $k \in 1, 2, 4$, and the ensemble encoder of all three ks.

4.2 Evaluation on Learned Representations

Unlike classic supervised learning where the model is evaluated based on its performance in recovering the same set of labels, our goal is not to evaluate the model for its ability to rank frames but to evaluate the learned representation on downstream tasks. First, we train a Resnet-18 encoder using a ranking triplet loss with VoxCeleb dataset, then we select the best encoder based on the ranking performance on the validation set (from VoxCeleb). We evaluate the representations generated by the encoder for facial action unit detection and expression recognition tasks on independent datasets by training a linear classifier, similar to [18] and [13].

The linear classifier consists of two layers: a batch-norm layer followed by a linear fully connected layer with no bias. We evaluate the methods on both embeddings learned from a single encoder and concatenated embeddings from ensemble encoders trained with ranking triplet loss.

Datasets: We evaluate facial action unit (AU) detection on three datasets including BP4D [37], DISFA [21], and EmotioNet [7]. We also evaluate our method on expression recognition on AffectNet [24]. We followed the same procedure as [18, 31] for BP4D and DISFA, evaluating the AU detections using a subject-independent 3-fold cross-validation. Emotionet and AffectNet are evaluated in similar fashion to [13], where the training set is split for training and validation, and an independent set is used for testing. The technical details of the prepossessing are available in the supplementary material.

Method	1	2	4	6	7	10	12	14	15	17	23	24	avg
Self-supervised													
TCAE [18]*	43.1	32.2	**44.4**	**75.1**	**70.5**	80.8	**85.5**	**61.8**	**34.7**	**58.5**	**37.2**	**48.7**	**56.1**
FAb-Net [13]*	**43.3**	**35.7**	41.6	72.9	63.0	75.9	83.5	57.7	26.5	48.2	33.6	42.4	52.0
TCAE (Re.) [18]	33.5	32.2	43.8	73.7	67.7	80.1	81.5	57.4	26.5	54.5	23.2	31.8	50.5
FAb-Net (Re.) [13]	33.4	24.8	41.0	73.5	66.2	78.8	84.7	57.9	21.2	55.7	26.8	37.9	50.2
Ranking k=1	35.2	25.5	30.2	71.3	69.6	**81.3**	83.3	59.1	30.3	56.1	27.0	33.4	50.2
Ranking Ensemble	42.3	24.3	44.1	71.8	67.8	77.6	83.3	61.2	31.6	51.6	29.8	38.6	52.0
Supervised													
AlexNet [2] *	40.3	39.0	41.7	62.8	54.2	75.1	78.1	44.7	32.9	47.3	27.3	40.1	48.6
DRML [39]*	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
EAC-Net [17]*	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
JAA-Net [31]*	47.2	44.0	54.9	77.5	74.6	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0

Table 1: F1-scores on BP4D dataset. * denotes values reported in original work and Re. indicates our reproduced results.

Binary cross entropy (BCE) loss is used for training a binary classifier for each AU. The explicit AU descriptions can be found in supplementary materials. Since AU labels are highly unbalanced, the loss for under-represented classes are weighted inversely proportional to their frequencies. The F1-score is used to evaluate BP4D and DISFA performances, the performance on EmotioNet and AffectNet is measured by Area under the ROC curve (AUC-ROC), similar to previous work [13, 25].

Comparison with baselines: We compare our method with the state-of-the-art self-supervised facial expression analysis methods, *i.e.*, TCAE [18] and FAb-Net [13]. For TCAE, we re-trained networks using the original code, available on GitHub[2], and we downloaded the available pre-trained model for FAb-Net. For both methods, we trained a linear classifier on the benchmark datasets with the exact same splits and hyperparameters. However, since the evaluation code is not released for TCAE and the trained encoders do not perfectly match, our reproduced performance scores are slightly lower than the published values for BP4D and DISFA datasets. To have a fair comparison, we compare our model to the reproduced performances for FAb-Net and TCAE.

Table 1 shows the results on BP4D dataset. Our method with a single encoder at time window $k = 1$ performs similar to TCAE and FAb-Net. The average F1-score of all three self-supervised methods are very close. However, our ensemble encoders model outperforms both reproduced results from TCAE and FAb-Net on average. The ensemble method performs better on subtle lip movements such as AU14 (dimpler), AU15 (lip corner depressor), AU23 (lip tightener) and AU24 (lip pressor).

Evaluations on DISFA are shown on Table 2. Like BP4D, DISFA contains videos. Our method with a single encoder outperforms both (reproduced) TCAE and FAb-Net in terms of average F1-score. For ensemble encoders, the result is considerably higher than the two existing self-supervised methods. Similarly, our method is superior for detecting lip movements like AU12 (Lip Corner Puller), AU25 (Lips part) and AU26 (Jaw Drop). One possible explanation is that since our network is trained on VoxCeleb2 datasets which include speaking, the encoder likely learned lip movements through leveraging available speaking behavior. Since neither TCAE nor FabNet consider temporal dependencies, and they focus

[2] https://github.com/mysee1989/TCAE

Method	1	2	4	6	9	12	25	26	avg
Self-supervised									
TCAE [18] *	15.1	15.2	**50.5**	48.7	23.3	**72.1**	**82.1**	**52.9**	**45.0**
FAb-Net [13] *	15.5	16.2	43.2	**50.4**	23.2	69.6	72.4	42.4	41.6
TCAE (Re.) [18]	24.8	25.5	37.3	34.7	**31.1**	59.6	58.1	25.2	37.0
FAb-Net (Re.) [13]	**27.5**	19.6	28.7	45.2	20.9	65.6	67.9	24.0	37.4
Ranking k = 1	10.8	20.7	43.3	37.6	12.2	68.7	62.9	46.2	37.8
Ranking Ensemble	18.7	**27.4**	35.1	33.6	20.7	67.5	68.0	43.8	39.4
Supervised									
DRML [39] *	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
EAC-Net [17] *	41.5	26.4	66.4	50.7	80.5	89.3	88.9	15.6	48.5
JAA-Net [31] *	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0

Table 2: F1-scores on DISFA dataset. * denotes values reported in original work and Re. indicates our reproduced results.

Method	1	2	4	5	6	9	12	17	20	25	26	avg
Self-supervised												
FAb-Net [13]*	73.4	71.8	75.3	67.8	90.4	78.8	91.9	72.4	74.5	83.7	**73.3**	77.6
TCAE [18]	74.1	72.6	79.8	74.3	91.4	83.2	93.4	73.7	75.5	83.7	72.8	79.5
FAb-Net (Re.) [13]	**75.1**	72.9	**82.1**	73.2	**92.2**	**86.4**	94.1	**76.4**	78.4	83.7	72.4	80.6
Ranking k = 1	68.1	71.4	78.5	76.2	91.5	80.0	94.7	71.8	75.4	84.0	69.3	78.3
Ranking Ensemble	70.7	**73.3**	80.5	**82.1**	92.1	**84.3**	**95.9**	73.4	**81.6**	**87.4**	72.2	**81.2**
Supervised												
VGG-Face [13]*	81.8	83.0	83.5	81.8	92.0	90.9	95.7	80.6	85.2	86.5	73.0	84.9
VGG-11 [13]*	74.7	77.2	85.8	83.7	93.8	89.7	97.5	78.3	86.9	96.4	81.5	86.0

Table 3: AUC on EmotioNet dataset. * denotes values reported in original work and Re. indicates our reproduced results.

on large pixel reconstruction losses instead of subtle movements, our method is better suited for dynamic facial AU detection. In addition, ranking triplet methods outperform traditional supervised methods like AlexNet [2] and DRML[16] in average F1-score, since those two methods are holistic and static. However, our method underperforms JAA-Net [31] and EAC-Net [17], where both supervised methods use facial landmarks to build patch-specific encoders. It is worth noting that all supervised results are directly taken from the original work. A major difference between our method and the supervised models is that we only train a single layer linear classifier on top of the encoder with frozen weights. This is a reason why the performance of self-supervised methods is lower than supervised ones, and that both baseline methods (TCAE, FAbNet) underperform the supervised models. We take this approach since the goal of our work is to evaluate the representations as opposed to targeting the final downstream tasks.

Results on EmotioNet are provided in Table 3. Unlike continuous recording in a restricted environment like BP4D and DISFA, EmotioNet database is collected in-the-wild. Our method with a single encoder is comparable to TCAE and performs slightly worse than FAb-Net in area under the ROC curve (AUC-ROC). The ensemble method outperforms both baselines, by improving the generalizability of the network for adopting discrete AU recognition. Furthermore, the ensemble method is not far behind supervised VGG descriptors, showing its generalization under a variaty of image domains. In addition to facial AU detection, the expression recognition results are shown in Table 4 for AffectNet. Both single and ensemble encoders are superior to the baselines, which further demonstrates the ability of the learned

Method	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt	avg.
Self-supervised									
FAb-Net [13] *	71.5	90.0	70.8	78.2	77.4	72.2	75.7	72.1	76.0
TCAE (Re.) [18]	66.8	81.9	65.5	70.6	70.8	63.6	70.3	68.5	70.0
FAb-Net (Re.) [13]	70.9	86.8	68.9	76.1	75.6	68.2	73.4	73.7	74.2
Ranking k = 1	70.6	90.0	70.8	76.7	78.2	73.6	75.2	74.2	76.2
Ranking Ensemble	**73.4**	**91.4**	**75.5**	**79.9**	**81.6**	**74.2**	**76.5**	**76.0**	**78.6**
Supervised									
VGG-Face [13]*	75.9	92.2	80.5	81.4	82.3	81.4	81.2	77.1	81.5
VGG-11 [13]*	–	–	–	–	–	–	–	–	82

Table 4: AUC on AffectNet dataset. * denotes values reported in original work and Re. indicates our reproduced results.

Method	BP4D	DISFA	EmotioNet	AffectNet
Triplet	48.1	29.8	74.4	70.2
N-pair-mc	41.5	18.5	62.8	59.3
Ranking k = 1	50.2	37.8	78.3	76.2
Ranking k = 2	49.9	**40.5**	78.0	75.5
Ranking k = 4	50.7	34.0	77.3	74.3
Ranking Ensemble	**52.0**	39.4	**81.2**	**78.6**

Table 5: Average F1-scores and AUCs on BP4D, DISFA, EmotioNet, AffectNet datasets

representations for facial expression analysis.

4.3 Ablation Study

In this section, we discuss various ablations and effects of different interval values k on the performance. We trained the encoder with three different values of $k = 1, 2, 4$, and compared their performance on AU detection and expression recognition. We also compare the performance of the ensemble of the three encoders versus each individual encoder. The simplest method involves using a single triplet loss as defined in Section 3.2, which we refer to as 'Triplet'. Instead of having a sequence of frames in a single loss, each training sample has only one triplet containing a fixed anchor, its adjacent frame as the positive and only one of the negative frames from our original sequence. Another simplification is directly using the exact N-pair-mc loss from the original paper which compares the anchor sample with all negative samples simultaneously.

Average performance on four datasets are shown in Table 5. The detailed individual AU scores are available in supplementary material. The results clearly demonstrate that the single-encoder models, regardless of time spacing k, outperform the simpler Triplet and N-pair-mc models across all datasets. This demonstrates the value of temporal dynamics in facial expression analysis. N-pair-mc's poor performance compared to the trivial triplet loss is surprising, this could be due to N-pair-mc's use of L2 penalty loss to regularize embeddings during training instead of normalization. This might result in instability of the numerical output affecting classification performance. Different values of k impact the results as well. Although $k = 1$ and $k = 2$ demonstrate similar results, the performance drops with $k = 4$ for the majority of the datasets. This indicates that the learned facial activities are short-term. The single encoder has the best overall performance at $k = 1$.

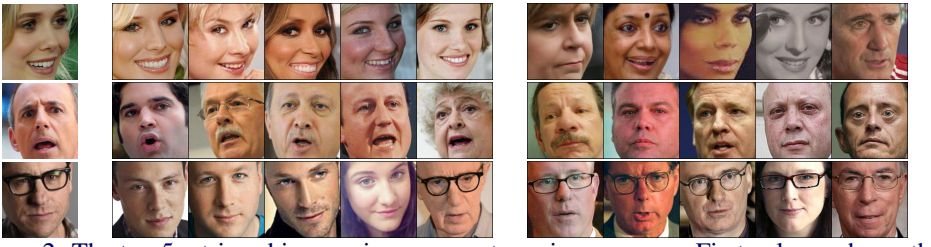

Figure 2: The top 5 retrieved images in response to an image query. First column shows the query image. The second column shows our results and the third column are FAb-Net's results. Query samples are intended to select different kinds of facial expressions and attributes (grin, surprise, contempt with eye-wear) to increase the diversity of queries.

4.4 Image Retrieval

We perform an image retrieval task on EmotioNet [7] to demonstrate the ability of the representations to capture expression similarity. We rank the frames based on the cosine similarity between their embeddings with a randomly selected query. The top five images with the highest cosine similarity to three queries are shown in the second column of Fig. 2. We also perform the same retrieval task using FAb-Net embeddings, displayed in the right column. The results demonstrate that our method retrieves better matches for facial expressions whereas FAb-Net tends to focus more on appearance and macro movements such as pose and appearance. For instance, the query on third row shows a man wearing glasses with closed lips and an activated buccinator muscle (AU14). Retrieved images using our model consist of similar facial expressions regardless of glasses. However, all images retrieved by FAb-Net include glasses, therefore showing dependence on appearance as well as expressions. More image retrieval examples are available in the supplementary material.

5 Conclusion

In this paper, we proposed a self-supervised learning method using temporal ranking triplet loss for facial AU recognition. By learning from inherent temporal consistency in videos, we achieved comparable or superior performance compared to the state-of-the-art self-supervised methods in facial AU detection and facial expression analysis. Through qualitative analysis, we also demonstrated the ability of the learned representations in capturing facial activities rather than appearance. With this work, we demonstrated the power of the self-supervised representation learning for generalizable facial expression analysis.

6 Acknowledgements

Research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. This research was also supported by Research on Azure Program of Microsoft.

References

[1] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gérard Medioni. Deep, landmark-free fame: Face alignment, modeling, and expression estimation. *International Journal of Computer Vision*, 127(6-7):930–956, 2019.

[2] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 25–32. IEEE, 2017.

[3] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Paul Ekman. Facial action coding system, 1977.

[6] Itir Onal Ertugrul, Jeffrey F Cohn, László A Jeni, Zheng Zhang, Lijun Yin, and Qiang Ji. Cross-domain au detection: Domains, learning approaches, and measures. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[7] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[9] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[10] 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*, pages 770–778, 2016.

[11] 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*, pages 4700–4708, 2017.

[12] Eric Jang, Coline Devin, Vincent Vanhoucke, and Sergey Levine. Grasp2vec: Learning object representations from self-supervised grasping. *arXiv preprint arXiv:1811.06964*, 2018.

[13] A Sophia Koepke, Olivia Wiles, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *BMVC*, page 302, 2018.

[14] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017.

[15] Gil Levi and Tal Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 503–510, 2015.

[16] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2017.

[17] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 103–110. IEEE, 2017.

[18] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10924–10933, 2019.

[19] Yen Khye Lim, Zukang Liao, Stavros Petridis, and Maja Pantic. Transfer learning for action unit recognition. *arXiv preprint arXiv:1807.07556*, 2018.

[20] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 2017.

[21] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[22] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[23] Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep learning from temporal coherence in video. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 737–744, 2009.

[24] 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*, 10(1):18–31, 2017.

[25] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[26] Itir Onal Ertugrul, Le Yang, Laszlo A Jeni, and Jeffrey F Cohn. D-pattnet: Dynamic patch-attentive deep network for action unit detection. *Frontiers in Computer Science*, 1:11, 2019.

[27] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[28] Magdalena Rychlowska, Rachael E Jack, Oliver GB Garrod, Philippe G Schyns, Jared D Martin, and Paula M Niedenthal. Functional smiles: Tools for love, sympathy, and war. *Psychological science*, 28(9):1259–1270, 2017.

[29] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering.

[31] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 705–720, 2018.

[32] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pages 1857–1865, 2016.

[33] Siyang Song, Enrique Sánchez-Lozano, Linlin Shen, Alan Johnston, and Michel Valstar. Inferring dynamic representations of facial actions from a still image. *arXiv preprint arXiv:1904.02382*, 2019.

[34] Raviteja Vemulapalli and Aseem Agarwala. A compact embedding for facial expression similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5683–5692, 2019.

[35] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M Robertson. Ranked list loss for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5207–5216, 2019.

[36] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

[37] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[38] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[39] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.