View-Invariant Compact Contrastive Learning for Facial Expression Recognition

Shuvendu Roy, Ali Etemad

Dept. ECE and Ingenuity Labs Research Institute
Queen's University, Kingston, Canada
{shuvendu.roy, ali.etemad}@queensu.ca

Abstract

We propose ViewFX, a novel view-invariant facial expression recognition framework based on contrastive learning. Using our newly proposed loss function, the model learns facial expression representations by clustering together different views of the same emotions for each subject to learn view-invariant features of expression. Next, class information are used in our model to form an external cluster of expression classes consisting of the previously formed view clusters. Lastly, to deal with the challenging subtle differences among facial expressions, we impose an extra learning step in the intermediate layers of the encoder to learn compact and de-correlated representations. We test the proposed framework on two public multi-view facial expression recognition datasets, KDEF and DDCF. The experiments demonstrate that our approach outperforms related works in the area and sets a new state-of-the-art for both datasets, while showing considerably less sensitivity to the amount of output labels used for training.

1 Introduction

Facial expressions are a crucial form of non-verbal communication, thus, facial expression recognition (FER) solutions can play a key role in human-computer interaction systems by allowing the system to understand and adapt to human reactions. Examples of such systems include health-care assistants (Tokuno et al. 2011), diving assistants (Leng, Lin, and Zanzi 2007), personal mood management systems (Thrasher et al. 2011; Sanchez-Cortes et al. 2013), emotion-aware multimedia and smart devices (Cho, Julier, and Bianchi-Berthouze 2019), and others. Nonetheless, FER remains a challenging task due to a number of issue such as the subtlety of facial actions that result in display of expressions.

In practical settings, it is highly likely for images of human faces to be captured from different angles and not always adhere to a clear and standard frontal view, creating a class of problem often referred to as 'multi-view' FER. The appearance of facial expressions from different angles can be quite different from one another, posing unique challenges due to large discrepancies between the extreme angles, e.g., profile vs. frontal. To this end, considerable efforts have been made to eliminate such view discrepancies and learn *view-invariant* representations for FER (Eleftheriadis, Rudovic, and Pantic 2014; Moore and Bowden 2011; Zhang et al. 2020).

A number of prior works have proposed multi-view solutions for FER (Mahesh et al. 2021; Liu et al. 2018, 2019). In (Liu et al. 2018), improvements were achieved for multi-view FER with a multi-channel pose-aware CNN (MPCNN). For recognizing the expressions and jointly learn pose at the same time, a hierarchical pose adaptive attention network was proposed in PhaNet (Liu et al. 2019). In this work, the attention module pays attention to view-invariant features at any angle, resulting in a more robust FER model. In (Roy and Etemad 2021), the authors used standard contrastive learning to perform multi-view FER, obtaining strong results.

To further expand on the problem statement, it has been well documented that facial expressions are determined by very subtle differences in the face (caused by small facial muscles) (Samal and Iyengar 1992). This results in higher than normal correlations between different classes (expressions) in the learned feature space. As a result, particular solutions capable of learning compact and de-correlated feature-spaces are required to achieve robust FER.

In this work, to tackle the problem of view-invariant FER, we propose a novel solution based on contrastive learning, which we name ViewFX. Our method uses multi-view information to bring different views of the same subject-expression close to each other in the embedding space, while pushing the clusters away from different expressions. This in turn will result in better and more robust FER, as well as reduced sensitivity to extreme angles. To achieve this objective, we propose a multi-view contrastive loss that forms clusters of different views of subject-expressions in the embedding space. The multi-view contrastive loss is combined with a supervised contrastive loss (Khosla et al. 2020), which forms external clusters consisting of the previously formed subject-expression subclusters. An additional term called compact loss is finally used to reduce the redundancy and correlations in the learned representation space, resulting in more compact expression-related embeddings. Successive to using the contrastive learning step to train an encoder, the model is fine-tuned on the downstream classification task. We rigorously evaluate our proposed method on two public multi-view FER datasets. Our experiments show that ViewFX achieves state-of-the-art results on both datasets, outperforming all the previous work in the area.

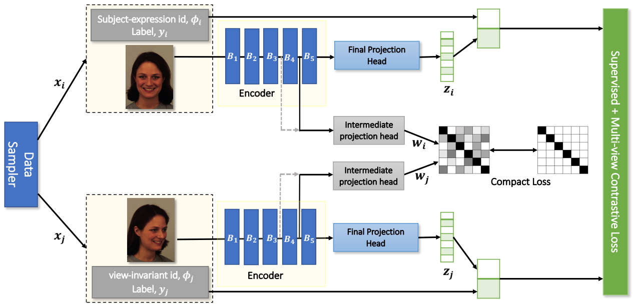

Figure 1: Illustration of the proposed training framework. Here, $B_1, \cdots, B_5$ represent Block$_1$ to Block$_5$ in the proposed encoder. Intermediate outputs from B_3 and B_4 feed the two individual instances of intermediate projection head.

2 Method

Assume dataset $\mathcal{D}$ consists of expressive faces $\{I\}_{\theta=1,\cdots,v}^{\phi=1,\cdots,p}$ and class labels (expression classes) $\{y\}^{\phi=1,\cdots,p}$, where ϕ denotes the subject-expression label for a total of p unique subject-expressions, and θ denotes the view from which each subject-expression is captured (a total of v views). Let $\{\theta_d, \cdots, \theta_q\}$ be specific views from which identifying I_y^x is challenging (e.g., due to sharp camera angles). To tackle this and effectively learn to classify I_y^x from these challenging views, we propose a novel contrastive framework, ViewFX, with a two-step training procedure. In the first step, we pre-train an encoder with a novel contrastive approach to learn ***view-invariant*** features from $\{I\}_{\theta=1,\cdots,v}^{\phi=1,\cdots,p}$ by clustering together representations corresponding to $\{\theta_1, \cdots, \theta_v\}$ for each subject-expression. Following this step, the encoder is fine-tuned for the down-stream task of classifying y^ϕ. In the next subsections, we describe the proposed loss and architecture of our contrastive framework.

The contrastive learning framework (Chen et al. 2020) learns useful features with positive and negative samples, where positive samples are different representations of I^i which are often generated by applying augmentations on the source image, and the negative samples are instances of any other image I^j, where $i \neq j$. A main flaw of this approach is that when it is followed by downstream classification, as pointed out in (Khosla et al. 2020), it may consider images of the same class as negative samples, hence situating them far apart in the embedding space. Consequently, (Khosla et al. 2020) proposed to use the class information to define the positive and negative samples to overcome this issue. Accordingly, if we modify the naive contrastive loss by bringing the images of the same class closer in the embedding space and moving them away from instances of other classes, a more effective and accurate embedding space will be achieved. This modification results in:

$$\mathcal{L}^{sup} = \sum_{i=1}^{2N} \left(\frac{-1}{2N_{y_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{y_i = y_j} \cdot \right. \\ \left. log \frac{exp(z_i \cdot z_j / \tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} exp(z_i \cdot z_k / \tau)} \right), \tag{1}$$

where, y_i is the class label of instance i and $2N_{y_i} - 1$ is the number of positive samples from class y_i.

In order to group together $\{I\}_{\theta=1,\cdots,v}^{\phi}$ for each ϕ in the embedding space to generate view-invariant representations, we propose a multi-view contrastive loss

$$\mathcal{L}^{view} = \sum_{i=1}^{2N} \left(\frac{-1}{2N_{\phi_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{\phi_i = \phi_j} \cdot \right. \\ \left. log \frac{exp(z_i \cdot z_j / \tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} exp(z_i \cdot z_k / \tau)} \right), \tag{2}$$

where, $2N_{\phi_i} - 1$ is the number of positive samples, i.e. different views and augmentations for a subject-expression label ϕ_i. $\mathcal{L}^{view}$ is derived from $\mathcal{L}^{sup}$ to facilitate multi-view representation learning by bringing together the embeddings of positive samples (different views with each subject-expression label), and moving them away from the negative samples (images of different subjects or expressions).

Facial expressions are often distinguished by very subtle differences in the embedding space. Consequently, it is a desired property for FER solutions to learn de-correlated and discriminative representations from the image. Inspired by (Bardes, Ponce, and LeCun 2021; Zbontar et al. 2021) where a new loss term has been proposed to de-correlate learned representations to avoid mode collapse, we introduce a loss term for the intermediate layers of our encoder (see Figure 1). The goal of this component is to generate ***compact*** representations by measuring the correlation between the embedding vectors of the augmented images and forcing them to become as close to identity as possible. Through our proposed use of this term in the intermediate layers of the encoder as opposed to the final layer, we force intermediate and lower level representations to become less redundant as opposed to the final embedding. In particular, we believe this benefits the model as expression-related representations are likely to be learned in the layers 'close' to the final layer, but not necessarily the last layer. Our experiments in Section 3.2 (ablation) further confirm that this design choice indeed boosts performance.

Let's represent a batch of images with two augmentations (Aug^A and Aug^B) as $\hat{I}^A$ and $\hat{I}^B$, each containing N images. The embeddings from the intermediate layers of the encoder are represented as w^A and w^B for for $\hat{I}^A$ and $\hat{I}^b$ respectively. Accordingly, the cross-correlation matrix between the two representations is calculated as $C_{ij} = \sum_{b=1}^{N} w_{b,i}^A w_{b,j}^B / (\sqrt{\sum_{b=1}^{N} (w_{b,i}^A)^2} \sqrt{\sum_{b=1}^{N} (w_{b,j}^B)^2})$, where b is the batch index, and i and j are the vectors' dimension index. The loss function is accordingly defined as:

$$\mathcal{L}^{comp} = \sum_i (1 - C_{ii})^2 + \alpha \sum_i \sum_{j \neq i} C_{ij}^2, \tag{3}$$

where the first term is called the 'invariance' term, which tries to equate the diagonal elements of the cross-correlation matrix to 1, forcing it to learn representations that are invariant to distortion/augmentations applied to the image. The second term is called 'redundancy reduction', and tries to equate the off-diagonal term to 0, effectively de-correlating the different vector components of the representation. Accordingly, α is the coefficient of the second term, manually set between 0 and 1.

The total loss function of our proposed method is a combination of the 3 loss function mentioned above, where the multi-view contrastive loss brings different views of same image closer in the embedding space creating smaller clusters of learned representations. The supervised contrastive loss brings samples of the same class together and pushes apart those of different classes, effectively forming an external cluster of the view-invariant clusters. The third loss term forces the intermediate embeddings to become more compact and less redundant to learn more robust expression representations. Accordingly, the overall loss is:

$$\mathcal{L}^{total} = \mathcal{L}^{sup} + \gamma * \mathcal{L}^{view} + \beta * (\mathcal{L}^{comp_1} + \mathcal{L}^{comp_2}), \quad (4)$$

where, γ and β are the coefficients of the multi-view and compact loss terms, while $\mathcal{L}^{comp_1}$ and $\mathcal{L}^{comp_2}$ are the losses applied to the two branches of the encoder (see Figure 1).

Training Pipeline. Figure 1 depicts the training pipeline of the proposed method. A data sampler samples a batch of images from the training set, and applies two augmentations yielding two sets of images x_i and x_j. The sampled images are then passed through the encoder and projection heads. The output of encoder goes into the projection head and outputs the projection embedding vector z_i and z_j, on which the supervised contrastive loss and multi-view contrastive loss function is applied. The encoder outputs two other intermediate representations that are passed through the intermediate projection head to generate w_i and w_j, on which the compact loss is applied.

Network Architecture As described above, the encoder of the proposed method generates one final output embedding and two intermediate embeddings. We have adopted the ResNet-50 (He et al. 2016) architecture as the encoder in the ViewFX pipeline, and added two branches to generate two intermediate outputs. These intermediate branches are followed by an intermediate projection head (shallow neural network with convolution, adaptive average pooling, and fully connected layers), which generate two intermediate embeddings on which the compact loss is applied to obtain $comp1$ and $comp2$. The encoder consists of 5 main blocks, where the two intermediate representations are generated from the output of the 3rd and 4th blocks. In addition to the two intermediate projection heads discussed above, a final projection head (a shallow neural network with only fully connected layers) is used after the last block, which generates the final output embedding and shown to be useful for leaning good representations of data (Chen et al. 2020; Khosla et al. 2020; Zbontar et al. 2021).

Augmentation Module. In the contrastive learning settings, an augmentation module is used to generate positive augmented samples from a given image. The augmentations used are random resize crop, random color distortion, random horizontal flip, random blurring, and random gray-scaling. The module applies random flip and gray-scaling with a probability of 0.5. For random cropping, we cropped the image by a factor of 0.2 to 1. For colour distortion, the brightness, hue, saturation, and contrast are randomly changed with a coefficient of 0.4 to 1. The module randomly picks the parameters above for all of these augmentations to generate each batch.

Training. As mentioned above, the training of the proposed method is a two step protocol, where in the first step the encoder is trained with the $\mathcal{L}^final$ loss, followed by a fine-tuning step. The projection head is useful for learning effective features from the data in the pre-training stage, which is no longer required for the final down-stream task. We therefore discard the projection head and add a final classification head in its to predict the output class probability. At this stage, the newly added classification head is fine-tuned with the categorical cross entropy loss as the encoder part of the model remains frozen.

We pre-train the encoder with the proposed total loss for 1000 epochs using an Adam optimizer and a learning rate of 1e-4. In this step, we use a cosine learning decay and weight decay of 1e-4. Once the pre-training step is done, we fine-tune the model for 25 epochs using a learning rate of 1e-4 and plateau learning rate decay with a decay factor of 0.5 and patience of 3. In the fine-tuning step, random resize crop and random horizontal flip are used in the augmentation module.

Implementation. The proposed method is implemented with PyTorch and trained on 4 NVIDIA V100 GPUs. For all the experiments, the image resolution is set to 224 × 224.

3 Experiments and Results

3.1 Datasets

KDEF (Lundqvist, Flykt, and Öhman 1998) is a multi-view dataset for facial expressions collected from 140 subjects. This dataset contains 7 classes where each subject-expression is captured from 5 different camera angles. The views are abbreviated as +90°: full right (FR), +45°: half right (HR), 0°: straight (S), -45°: half left (HL), and -90°: full left (FL). **DDCF** (Dalrymple, Gomez, and Duchaine 2013) is another multi-view facial expression dataset which is collected from 80 subjects. The dataset is captured from 5 different camera angles: +60°: full right (FR)), +30°: half right (HR), 0°: straight (S), -30°: half left (HL), and -60°: full left (FL). The DDCF dataset contains 8 facial expression classes.

3.2 Performance

Results and Comparison to Other Methods. We perform a number of experiments to fully evaluate the performance of ViewFX and the role of different components and parameters. Table 1 presents the accuracies of our method in comparison with prior works in the area, averaged over all the angles. It should be noted that the majority of other papers don't report F1. However, we do for completeness, we report here that the F1 scores for KDEF and DDCF are 96.97±0.59 and 97.30±0.64 respectively. We observe that in comparison against the state-of-the-art, ViewFX achieves more than 2% improvement for the DDCF dataset and 2.5% improvement for the KDEF dataset.

Table 1: Performance and comparison to previous works on KDEF and DDCF datasets.

Dataset	Method	Acc. ± SD
KDEF	SVM (Moore and Bowden 2011)	70.5±1.2
	SURF (Rao et al. 2015)	74.05±0.9
	TLCNN (Zhou and Shi 2017)	86.43±1.0
	PhaNet (Liu et al. 2019)	86.5
	MPCNN (Liu et al. 2018)	86.9±0.6
	RBFNN (Mahesh et al. 2021)	88.87
	CL-MEx (Roy and Etemad 2021)	94.64±0.92
	ViewFX (Ours)	**97.15±0.46**
DDCF	LBP (Khan et al. 2019)	82.3
	SVM (Albu et al. 2015)	91.27
	RBFNN (Albu et al. 2015)	91.80
	CL-MEx (Roy and Etemad 2021)	95.26±0.84
	ViewFX (Ours)	**97.34±0.54**

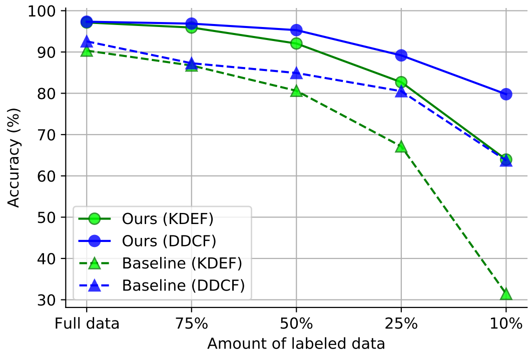

Figure 2: Performance for different amounts of output labels used in training.

Sensitivity to Output Labels. Next, we evaluate the sensitivity of our method against the amount of labeled data used. Figure 2 presents the performance of our method in comparison to the baseline (cross-entropy loss) when the amount of labeled data is reduced. We observe that ViewFX maintains a stable performance far better than the baseline when the amount of output labels during training is reduced to 75%, 50%, 25%, and even 10%. For instance, in the extreme case of using only 10% of the labels, the drops in accuracy are approximately 12% and 30% for our method in KDEF and DDCF, in comparison to the baseline's respective 25% and 60% drops, highlighting considerable less sensitivity to ground-truth labels by our method.

Ablation Experiments. To understand the impacts of different components of the proposed framework, we conduct ablation studies. Table 2 shows the results when different ablations and variations of the total loss are used to train the model. As we can see, the proposed loss has significant improvement over the baseline (cross-entropy loss) with nearly 7% and 5% improvements for KDEF and DDCF datasets respectively. Next, systematically remove each loss term from the framework to evaluate its impact.

Table 2: Ablation experiments on the different components of the proposed framework. Compact(inter.) denotes the proposed compact loss used in the intermediate layers, while Compact(final) denotes the compact loss used for the final layer. The results are presented for when 100% and 50% output labels are used for training (in 100% / 50% format).

Loss	KDEF	DDCF
Cross-Entropy (baseline)	90.34 / 80.61	92.58 / 84.92
Supervised Cont. + Compact(inter.)	96.54 / 90.02	96.55 / 94.20
Multi-view Cont. + Compact(inter.)	94.80 / 84.93	94.72 / 90.44
Supervised Cont. + Compact(final)	96.33 / 88.75	95.81 / 90.25
Multi-view Cont. + Compact(final)	94.68 / 81.53	94.62 / 86.98
Supervised Cont. + Multi-view Cont.	96.95 / 90.84	96.87 / 93.26
Total loss	**97.15 / 92.06**	**97.34 / 95.30**

First, we remove the multi-view contrastive component from our framework, followed by the removal of the supervised contrastive component from the total loss function, where both cases, we observe a drop in performance. Next, we repeat this experiment with the exception of using the compact loss for training the final projection head instead of intermediate projection head. This experiment validates the choice of using the compact loss for the intermediate layers as opposed to the final projection head. Lastly, we remove the compact loss term altogether and only use the supervised and multi-view contrastive losses, demonstrating the positive impact of the compact loss. We perform these experiments twice, once with all the labels used in training, and once when only 50% of the output labels are used. This experiment further highlights the added value of our method when smaller amounts of labels are available for training.

4 Conclusion

This paper introduces a contrastive learning framework for view-invariant FER. Our method uses a novel loss function that consists of three specific terms: the first term learns view-invariant representations for facial expressions; the second term learns supervised expression class information; and the third term forces the embedding representations to become compact by learning and de-correlating subtle differences between expressions. We test our proposed framework on two multi-view datasets, KDEF and DDCF, where we achieve state-of-the-art performances when a single-view image is used at inference. The model exhibits considerably less reliance on the amount of output labels used for training. For future work, the proposed method can be used for other multi-view recognition objectives, for example multi-view face recognition or multi-view object recognition.

References

Albu, F.; Hagiescu, D.; Vladutu, L.; and Puica, M.-A. 2015. Neural network approaches for children's emotion recognition in intelligent learning applications. In *7th Annu Int Conf Educ New Learn Technol Barcelona*. 4

Bardes, A.; Ponce, J.; and LeCun, Y. 2021. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*. 2

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. 2, 3

Cho, Y.; Julier, S. J.; and Bianchi-Berthouze, N. 2019. Instant stress: detection of perceived mental stress through smartphone photoplethysmography and thermal imaging. *JMIR Mental Health*, 6(4). 1

Dalrymple, K. A.; Gomez, J.; and Duchaine, B. 2013. The Dartmouth Database of Children's Faces: Acquisition and validation of a new face stimulus set. *PloS One*, 8(11): e79131. 3

Eleftheriadis, S.; Rudovic, O.; and Pantic, M. 2014. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on image processing*, 24(1): 189–204. 1

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778. 3

Khan, R. A.; Meyer, A.; Konik, H.; and Bouakaz, S. 2019. Saliency-based framework for facial expression recognition. *Frontiers of Computer Science*, 13(1): 183–198. 4

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. *Advances in Neural Information Processing Systems*, 33. 1, 2, 3

Leng, H.; Lin, Y.; and Zanzi, L. 2007. An experimental study on physiological parameters toward driver emotion recognition. In *International Conference on Ergonomics and Health Aspects of Work with Computers*, 237–246. 1

Liu, Y.; Peng, J.; Zeng, J.; and Shan, S. 2019. Pose-adaptive Hierarchical Attention Network for Facial Expression Recognition. *arXiv preprint arXiv:1905.10059*. 1, 4

Liu, Y.; Zeng, J.; Shan, S.; and Zheng, Z. 2018. Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*, 458–465. 1, 4

Lundqvist, D.; Flykt, A.; and Öhman, A. 1998. The Karolinska Directed Emotional Faces (KDEF). *CD ROM from Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet*, 91(630): 2–2. 3

Mahesh, V. G.; Chen, C.; Rajangam, V.; Raj, A. N. J.; and Krishnan, P. T. 2021. Shape and Texture Aware Facial Expression Recognition Using Spatial Pyramid Zernike Moments and Law's Textures Feature Set. *IEEE Access*, 9: 52509–52522. 1, 4

Moore, S.; and Bowden, R. 2011. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4): 541–558. 1, 4

Rao, Q.; Qu, X.; Mao, Q.; and Zhan, Y. 2015. Multi-pose facial expression recognition based on SURF boosting. In *IEEE International Conference on Affective Computing and Intelligent Interaction*, 630–635. 4

Roy, S.; and Etemad, A. 2021. Self-supervised Contrastive Learning of Multi-view Facial Expressions. *arXiv preprint arXiv:2108.06723*. 1, 4

Samal, A.; and Iyengar, P. A. 1992. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern recognition*, 25(1): 65–77. 1

Sanchez-Cortes, D.; Biel, J.-I.; Kumano, S.; Yamato, J.; Otsuka, K.; and Gatica-Perez, D. 2013. Inferring mood in ubiquitous conversational video. In *12th International Conference on Mobile and Ubiquitous Multimedia*, 1–9. 1

Thrasher, M.; Van der Zwaag, M. D.; Bianchi-Berthouze, N.; and Westerink, J. H. 2011. Mood recognition based on upper body posture and movement features. In *International Conference on Affective Computing and Intelligent Interaction*, 377–386. 1

Tokuno, S.; Tsumatori, G.; Shono, S.; Takei, E.; Yamamoto, T.; Suzuki, G.; Mituyoshi, S.; and Shimura, M. 2011. Usage of emotion recognition in military health care. In *Defense Science Research Conference and Expo*, 1–5. 1

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*. 2, 3

Zhang, F.; Zhang, T.; Mao, Q.; and Xu, C. 2020. Geometry guided pose-invariant facial expression recognition. *IEEE Transactions on Image Processing*, 29: 4445–4460. 1

Zhou, Y.; and Shi, B. E. 2017. Action unit selective feature maps in deep networks for facial expression recognition. In *IEEE International Joint Conference on Neural Networks*, 2031–2038. 4