Exploring Expression-related Self-supervised Learning and Spatial Reserve Pooling for Affective Behaviour Analysis

Fanglei Xue[1*] Yifan Sun[2] Yi Yang[3†]
[1] University of Technology Sydney [2] Baidu Inc. [3] Zhejiang University
xuefanglei19@mails.ucas.ac.cn, sunyf15@tsinghua.org.cn, yangyics@zju.edu.cn

Abstract

Self-supervised learning (SSL) methods have gained attention for reducing dependence on labeled data. However, SSL methods are less investigated for facial expression recognition (FER), which requires expensive expression annotation, especially for large-scale video databases. In this paper, we explore an expression-related self-supervised learning (SSL) method called ContraWarping to perform expression classification in the 5th Affective Behavior Analysis in-the-wild (ABAW) competition. We also conduct a new spatial reserve pooling module to utilize all facial details for expression recognition. By evaluating on the Aff-Wild2 dataset, we demonstrate that ContraWarping outperforms existing supervised methods and other general SSL methods with only 0.7M trainable parameters and shows great application potential in the affective analysis area. Codes have been released at https://github.com/youqingxiaozhua/ABAW5.

1. Introduction

Affective computing aims to recognize expressions from static images or videos automatically. With affective computing, people could build applications in society analysis, human-computer interaction systems, driver fatigue monitoring, and so on. For the past few years, many methods [9, 26, 31, 35, 38, 46, 47] have been proposed to recognize expressions. However, these methods all rely on precise human annotations to learn. Although some of them [34, 46, 47] could learn from noisy labels, they can not learn from unlabeled data. Unfortunately, expressions are subjective and subtle, making annotation a large-scale expression database very expensive and limiting the scale of current databases.

Recently, some researchers proposed some self-supervised learning methods to learn from unlabeled data.

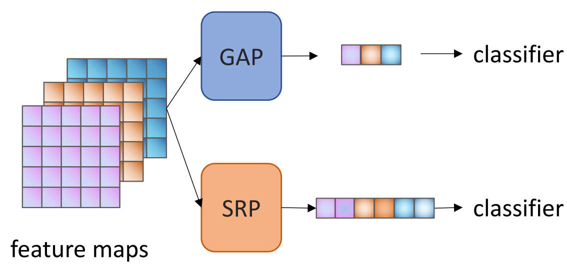

Figure 1. The traditional global average pooling (GAP) method pools multiple feature maps among the spatial dimension, resulting in a single scalar for each feature map. This causes the loss of spatial information. However, the proposed spatial reserve pooling (SRP) module could address this issue by preserving spatial features to better utilize features from different facial areas.

Contrastive learning-based methods (such as SimCLR [2], MoCo [11], BYOL [7], *etc*.) learn image features from different views of the same images with a Siamese network. Differently, MAE [10] try to reconstruct a masked image to learn semantic features. Some works also adopt these ideas for face tasks. SSPL [33] learns the spatial-semantic relationship of face images by correct rotated patches, face parsing, and area classification tasks. He *et al*. [13] try to benefit the face recognition task by adopting a 3D reconstruction task. TCAE [28] and FaceCycle [43] learn face representation by disentangling pose, expression, and identity features from each other. Most recently, a contrastive learning method, ContraWarping [36], is proposed to learn expression-related features by directly simulating muscle movements. All these methods demonstrate their effectiveness in static image databases [1, 23, 26].

Aff-wild2 [15–24, 40] is a large-scale video database for ABAW competitions. It annotated 548 videos, around 2.7M frames, into eight pre-defined categories: anger, disgust, fear, happiness, sadness, surprise, neutral, and others. Thanks to the release of this database, we conduct experiments to explore the effectiveness of ContraWarping on this in-the-wild video database. By directly fine-tuning a part

*Work was down when Fanglei Xue was an intern at Baidu Research.
†Corresponding author.

Figure 2. The attention maps from the backbone. As we can see, different facial areas play different roles in recognition. However, the global average pooling (GAP) directly takes a mean among spatial features which lack spatial information. To reserve the spatial information of the whole face, we propose spatial reserve pooling (SRP) to replace the traditional GAP in recognition models.

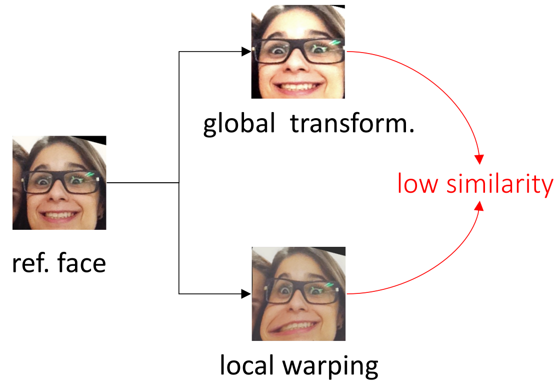

Figure 3. The concept illustration of ContraWarping [36] method. Given a reference face, ContraWarping generates two faces with global transformations and the proposed random local warping, respectively. The prior transformation does not change the expression while the second warping moves the facial muscles and changes the facial expression. By pushing these two faces away in the feature space, ContraWarping could learn expression-related features. (ContraWarping also pulls faces with different global transformations close, this branch is not illustrated for simplicity.)

of the pre-trained weights from ContraWarping, we demonstrate that recent SSL methods could extract more informative features than face recognition supervised counterparts. And the expression-related method, ContraWarping, performs better than other general SSL methods, indicating a great potential in expression recognition tasks.

On the other hand, current image classification methods always adopt a global average pooling (GAP) module to pool the 3D feature maps to a vector. In this process, the features on the same channel are averaged along the spatial dimension (both height and width), which loses the spatial information. We visualize the attention map of the feature maps from the backbone in Fig. 2. As we can see, the model pays attention to a large area among the faces, mainly including areas around eyebrows, eyes, nose, and mouth. Intuitively, the model needs both semantic and these spatial features to recognize the expression. To retain spatial information, we propose a spatial-reserved pooling (SRP) module to replace the traditional GAP. Specifically, two convolutional layers are utilized to reduce the channel and spatial dimensions. After that, the features are flattened instead of global pooling to reserve all information.

Combining with the expression-related SSL method ContraWarping and the new proposed SRP module, we get the performance of 37.57% f1-score on the validation set of Expression (Expr) Classification Challenge with a Res-50 backbone, significantly outperforming the supervised one. And without any temporal information, our method ranked 6th on the test set of ABAW5 with only **0.7M** trainable parameters.

2. Related Works

Many inspirational methods have been proposed in previous ABAW competitions. We investigate some expression classification methods and multi-task learning methods which, including the Expr task.

Zhang *et al*. [41] ensemble multiple 2D backbones to extract features for every single frame and concatenate these features to a temporal encoder to explore temporal features. By combining regression layers and classification layers, it learns from multi-task annotation and ranks first in the ABAW4 challenge. It also used MAE pre-trained weights to enhance its performance. Li *et al*. [27] also use MAE pre-trained weights combined with AffectNet supervised pre-trained weights and ranked 2nd in ABAW4. Zhang *et al*. [45] proposed a transformer-based fusion module to fuse multi-modality features from audio, image, and word information. Jeong *et al*. [14] extended the DAN model and achieved 2nd in ABAW3. Xue *et al*. [37] utilized a coarse-to-fine cascade network with a temporal smoothing strategy and ranked 3rd in ABAW3. Zhang *et al*. [42] found that AU, VA, and Expr representations are intrinsically associated with each other and proposed a streaming network for multi-task learning.

3. Method

Since this paper focuses on exploring the efficiencies of different self-supervised learning methods, we adopt a simple framework to directly perform frame-wise classification. Before introducing the architecture and implementation details, we first introduce some preliminaries of self-supervised learning methods in the facial area.

3.1. Preliminry

Recently, self-supervised learning methods have raised wide attention to learning from unlabelled data directly, giving a new solution to address the expensive annotation problem of FER databases. Different from supervised learning from human annotations, these methods push the model to solve a no annotation needed pretext task to learn representations, for example: predicting relative patches [5, 10], image inpainting [30], solving jigsaw [29], contrastive learning [8], masked image model [10], and so on. For example, various works [27,41] in ABAW competitions have adopted MAE to pre-train their models on a combination of numerous facial recognition databases and achieve promising performance. However, these pretext tasks are still designed for the common image classification task, which aims to recognize the species of foreground objects. It is less efficient to extract expression-related features by directly applying these methods to FER.

To bring expression information to the pretext task, CRS-CONT [25] adopts coarse-grained expression labels in the pre-training stage. Although coarse labels are more accessible to collect than fine-grained labels, they still need to label a great number of images. Recently, ContraWarping [36] was proposed to address this issue. It proposed a local warping method to simulate muscle movements and change the original expression without any human annotation. By pushing warped faces away in the feature space, it could learn expression-related features in a self-supervised learning manner. Fig. 3 illustrate the contrastive concept of ContraWarping.

3.2. Architecture

3.2.1 Overview

Empowered by this expression-related self-supervised learning method, models could learn to distinguish muscle movements and extract abundant expression features. Thus, we adopted a simple pipeline to investigate the capacity of ContraWarping on the Aff-Wild2 dataset. As illustrated in Fig. 4, the facial image (denoted as I) was firstly extracted by the backbone:

$$f_{map} = B(I) \tag{1}$$

where B denotes the backbone network, and f_{map} denotes the extracted feature maps. The feature maps are further aggregated by our proposed Spatial Reserve Pooling module to reduce dimensions and reserve spatial information. The per-class scores are calculated by a fully-connected layer:

$$scores = Softmax(FC(SRP(f_{map}))) \tag{2}$$

The model is trained with the cross-entropy loss, which can be formulated as:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} log(haty_{ij}) \tag{3}$$

where N is the number of batch samples, C is the number of predicted classes (eight for Aff-Wild2), and y_{ij}, $haty_{ij}$ represent the gound-truth label and predicted scores, respectively.

3.2.2 Backbone

Following [41], we simply adopt 2D backbones (such ResNet [12], ViT [6], *etc.*) to extract features from every frames. We do not use temporal features for simplicity and mainly focus on the evaluation of the effectiveness of SSL methods.

Since the ContraWarping adopted Res-18 and Res-50 as backbones, we utilized these two pre-trained backbones to finetune on the Aff-Wild2 dataset. As illustrated in Fig. 4, the backbone typically consists of four stages: the shallow stages are corresponding to extracting low-level features, such as lines and shapes while the deep stages are mainly focused on extracting abstract semantic features based on shallow layers' output.

Since the backbone is pre-trained with ContraWarping methods on a large number of face images, it has learned to distinguish muscle movements among faces and could extract informative and expression-related features. Experiments on RAF-DB also demonstrate good linear evaluation performance which freezes the whole backbone. To better utilize the capability of pre-training and adopt the pre-trained method on this large-scale video dataset, we freeze the first three stages (denoted as a snow mark in Fig. 4) and only finetune the last stage of the backbone to adapt semantic features to the downstream database.

3.2.3 Spatial Reserve Pooling

Multiple feature maps are extracted from the backbone model. As illustrated in Fig. 5, given a face image with shape 224×224, every feature map has a shape of 7×7. This indicates that every pixel in the feature map represents a semantic feature of a small region of the original input face image. As we have illustrated in 2, the model may need sufficient information from multiple face areas to distinguish one expression category from another. In other words, the spatial information in the feature maps is essential for facial expression recognition.

However, Traditional recognition models typically adopt a Global Average Pooling (GAP) between the backbone and

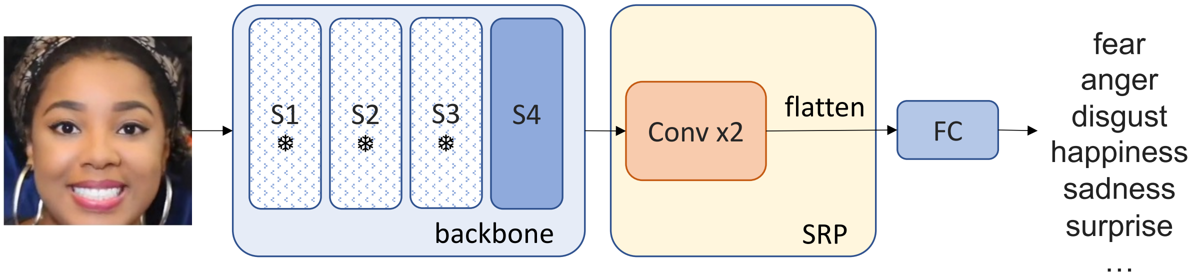

Figure 4. The illustration of the pipeline of our method. The image is first passed through a backbone to extract feature maps. These feature maps are then input into the proposed SRP module, which pools them to reduce dimensions while preserving spatial features. Finally, a single fully-connected layer is used to classify the pooled features into eight expression categories. The backbone is typically composed of four stages, denoted as S1-S4. During training, the first three stages of the backbone are frozen (indicated by a snow icon) to fully utilize the capacity of SSL pretraining.

the classifier. GAP takes an average among the spatial channel to reduce dimensions to satisfy the required shape of the classifier. This operation causes the loss of spatial information and forces the model to embed spatial information into channel dimensions. This relies on a large receptive field, which is another shortcoming of convolutional networks, making the model harder to train.

To address this issue, we propose Spatial Reserve Pooling to replace the traditional Global Average Pooling. As illustrated in Fig. 5, a convolutional layer is first used to aggregate spatial features among different feature maps and reduce the number of feature maps to a fixed number (e.g., 256). This can further increase the receptive field of the model and reduce the computation cost of the following models. After this, another convolutional layer with a stride of two is used to shrink the spatial dimension. After these two convolutional layers, the original feature map (in the shape of $7 \times 7 \times C$) has been squeezed to $3 \times 3 \times 256$. The features are then flattened and fed into the following classifier. In this manner, spatial information from all face areas is preserved, and the receptive field of the model is further enlarged to the whole face. On the other hand, the computation cost is not increased too much. We strike a good balance between accuracy and computation cost.

3.3. Implementation

3.3.1 Dataset

For this ABAW challenge, Kollias *et al*. collected a large-scale video database named Aff-Wild2. It consists of 548 videos and was labeled frame-by-frame. For the expression classification challenge, every frame in the video is annotated with one of eight pre-defined expression categories: anger, disgust, fear, happiness, sadness, surprise, neutral, and others.

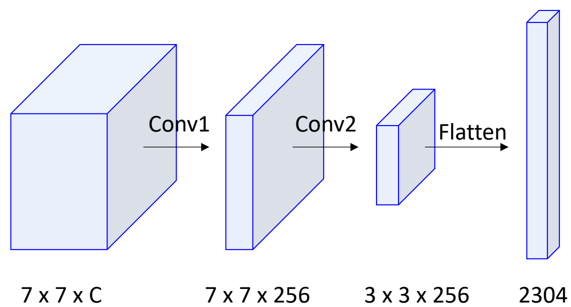

Figure 5. Illustration of the structure of our proposed special reserve pooling (SRP) module. C denotes the number of the feature map, *e.g.* 512 for a Res-18 backbone. The feature maps have a 7×7 spatial shape which indicates different expression features in different face areas. Our proposed SRP module aggregates these spatial features and reduces its dimension by two convolutional layers. Then the features are flattened to reserve expression features from all facial areas.

3.3.2 Metrics

The average f1 Score across all eight categories on the validation set is measured as a performance assessment.

$$P = \sum_{i=1}^{8} F_1/8 \tag{4}$$

where F_1 denotes f1-score, is calcuated by:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{5}$$

Pre-trained	F1-score
Sup. MS1M	27.71
SimSiam [3]	25.40
BYOL [7]	31.74
ContraWarping [36]	**34.85**

Table 1. Performance comparison with different initialization of the Res-18 backbone.

3.3.3 Experiment Setup

We adopt random cropping and horizontal flip for data augmentation to prevent over-fitting. The model is fine-tuned with the SGD optimizer for 8000 iters. The learning rate is set to 5e-3 with a cosine decay. The batch size is set to 128. Since the adjacent frames in the video are very similar, we randomly sample one frame of every ten frames for training.

By default, the Res-18 [12] network without the last classifier is adopted as the backbone for ablation studies. It takes about 10 minutes to train our method with two NVIDIA V100 GPUs. The Res-50 is utilized to generate the final results on the test set. The hidden dimension of two down-sampling convolution layers is set to 256.

4. Experiments

4.1. Ablation Studies

Comparison with different SSL methods. SimSiam [3] and BYOL [7] are two recently proposed SSL methods that utilize contrastive learning to learn informative features from unlabeled data. We compare them with supervised training on MS1M, a large-scale facial recognition database. The results are illustrated on Tab. 1. The SimSiam performs the worst since it does not rely on any supervised information during training, and it can not extract enough discriminative features for recognition. The supervised pre-training performs better than SimSiam but performs inferior compared with BYOL and ContraWarping. This indicates that the supervised pre-training could make the model convergence well so it avoids the model learning from scratch in downstream tasks. However, the face recognition task aims to recognize the identity of the face, which makes the model learns to suppress expression features. This task mismatch limits the performance of supervised ones. ContraWarping introduces expression-related contrastive information to the pre-training period by simulating muscle movements to change the expression. Experiment results also proved that expression-related pre-training could benefit downstream training.

Comparison with GAP with our proposed SRP. The GAP is widely used in image classification methods to reduce the dimension of the feature maps. However, it lost spatial information when performing the global average. To address this problem, we propose a spatial reserve pooling (SRP) module to combine both semantic and spatial information for recognition. To evaluate the effectiveness of the new proposed SRP, we conduct experiments with Res-18. As illustrated in Tab. 2, our proposed SRP increases the f1-score from 30.92% to 34.85%, indicating that spatial information is crucial for expression recognition and our SRP is efficient at delivering spatial information.

Pooling Method	F1-score
GAP	30.92
SRP	**34.85**

Table 2. Performance comparison with GAP and our proposed SRP.

Sample Method	F1-score
1/1	32.23
1/10	**34.85**

Table 3. Performance comparison with global average pooling and our proposed spatial reserve pooling.

Comparison with different sample methods for training. The expression changing in videos is continuous. However, we find that adjacent frames are very similar. There are about 2.7 million frames in the Aff-wild2 database. Using all frames for training is computationally expensive and inefficient. To reduce the computation cost and retrain the diversity of training frames as much as possible, we choose every one frame for training for every ten frames (denoted as 1/10). Specifically, we randomly set an offset value (denoted as j), and the frames are only selected if its frame id (i) is exactly divisible by j. The j is randomly selected for every epoch to retrain diversity.

To evaluate the effectiveness of this sampling method, we compare it with the default no-sampling method (denoted as 1/1) in Tab. 3. As we can see, the 1/10 sample strategy performs similarly (34.85%) and is one point better than the traditional no-sampling strategy (32.23%). This may be because our training period is short to prevent overfitting but indicates that our 1/10 sampling is efficient.

Determine the best freeze number of the backbone. Considering the backbone is already pre-trained with expression-related tasks, we try to freeze shadow stages of the backbone to keep the model's ability to extract expression features. This is also beneficial to preventing overfitting. As more parameters are frozen, fewer parameters are left to adapt to the target database, we conduct experiments on the validation set to determine the best freeze number.

Freeze number	F1-score
0	33.24
1	29.90
2	30.60
3	**34.85**
4	32.65

Table 4. Performance comparison with global average pooling and our proposed spatial reserve pooling.

Backbone	Pre-trained	F1-score
IR-50 [4]	Sup. MS1M	30.78
APViT [39]	Sup. MS1M	35.48
APViT [39]	Sup. RAF-DB	35.63
Res-18 [12]	ContraWarping	34.85
Res-50 [12]	ContraWarping	**37.57**

Table 5. Results with different backbones and pre-trained weights. Sup. indicates supervised pre-training with manually annotated labels.

As shown in Tab. 4, when finetuning the whole model (zero stage is frozen), it achieves 33.24% f1-score. But as we freeze more stages, the performance first decay to 39.90% and gradually increase to 34.85% when freezing the first three stages. The model could also achieve 32.65% on the f1-score without tuning the whole backbone, indicating the effectiveness of extracted features by ContraWarping.

Comparison under multiple backbones and pre-trained methods. To investigate the effectiveness of ContraWarping on this in-the-wild video database, we conduct experiments with several backbones and pre-trained weights on the validation set of ABAW5. As illustrated in Tab. 5, models with more parameters are not always better. APViT [39] is a recently proposed state-of-the-art method that combines both CNN and ViT for feature extraction. It boosts IR-50 from 30.78 to 35.48. However, it fails to outperform Res-50 with ContraWarping pre-trained, which achieves 37.57 on the validation set. The ContraWarping could increase the performance significantly. Even a simple Res-18 could outperform IR-50 with 34.85, indicating that ContraWarping pre-training is more suitable for expression analysis.

4.2. Results on the Test Set

We illustrate the performance of participants on the test set in Fig. 6. Our simple strategy achieves an f1-score of 0.3218, ranked sixth on the leaderboard.

The champion of the ABAW5 track [44] achieved an f1-score of 0.4121. They pre-trained the Masked Autoencoder (MAE) model on various large-scale facial image datasets

Team	Rank	F1-score
Netease Fuxi Virtual Human [44]	# 1	0.4121
SituTech	# 2	0.4072
CtyunAI [49]	# 3	0.3532
HFUT-MAC [48]	# 4	0.3337
HSE-NN-SberAI [32]	# 5	0.3292
Ours	# 6	0.3218

Table 6. Results with different backbones and pre-trained weights. Sup. indicates supervised pre-training with manually annotated labels.

and used it as a visual feature extractor. Their model also consists of a temporal and multi-modal fusion to leverage temporal and multi-modal information from videos. It is worth noting that they relied on a crowdsourcing platform to check and remove incorrect images in the processing progress. The second-place team [49] achieved a very competitive performance on the test set with an f1-score of 0.4072. The third team [49] also combined audio and image information as well as temporal features. Different from the champion strategy, the visual and audio features were first input into their respective temporal modules and then concatenated while the champion team first concatenated multi-modal features and passed them to temporal modules. The ranked fourth team [48] adopted a large number of state-of-the-art methods to extract visual features and proposed an affine module to align different features. The rank fifth team [32] ensemble multiple models from the EmotiEffNet family and achieved an f1-score of 0.3292.

The first-ranked method used multi-modality features, temporal features, model ensembles, and output smoothing strategies to improve performance. The third and fourth-ranked methods also used multi-modality and temporal features. The fifth-ranked method used model ensembles and output smoothing. Unlike these methods that aimed to improve performance, our goal was to investigate the effectiveness of expression-related SSL methods. We did not use the above-mentioned techniques and directly predicted every frame. Even so, we achieved an F1-score of 0.3218 on the test set, indicating the effectiveness of the ContraWarping pre-training method. Combining with multi-modality features and other good designs could also benefit performance.

5. Conclusion

In this paper, we adopt a simple pipeline to evaluate the effectiveness of ContraWarping, a self-supervised learning method for affective analysis on Aff-Wild2. The ContraWarping could learn expression-related features from unlabeled data by simulating muscle movements and could adapt well to downstream databases even with the first three stages frozen. Experiments on Aff-Wild2 indicate that models initialized with ContraWarping pre-trained weights could extract more informative features and performs better than supervised ones.

References

[1] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, 2016. 1

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*. arXiv, 2020. 1

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 5

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 6

[5] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised Learning. In *NeurIPS*, 2020. 1, 5

[8] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pages 1735–1742. IEEE, 2006. 3

[9] SL Happy and Aurobinda Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing*, 6(1):1–12, 2014. 1

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, pages 16000–16009, 2022. 1, 3

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 6687–6696, 2020. 1

[12] 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. 3, 5, 6

[13] Mingjie He, Jie Zhang, Shiguang Shan, and Xilin Chen. Enhancing Face Recognition With Self-Supervised 3D Reconstruction. In *CVPR*, pages 4062–4071, 2022. 1

[14] Jae-Yeop Jeong, Yeong-Gi Hong, Daun Kim, Yuchul Jung, and Jin-Woo Jeong. Facial expression recognition based on multi-head cross attention network. *arXiv preprint arXiv:2203.13235*, 2022. 2

[15] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 1

[16] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023. 1

[17] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020. 1

[18] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 1

[19] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 1

[20] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. *arXiv preprint arXiv:2303.01498*, 2023. 1

[21] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019. 1

[22] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1

[23] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 1

[24] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 1

[25] Hangyu Li, Nannan Wang, Xi Yang, and Xinbo Gao. CRS-CONT: A Well-Trained General Encoder for Facial Expression Analysis. *IEEE Transactions on Image Processing*, 31:4637–4650, 2022. 3

[26] Shan Li, Weihong Deng, and JunPing Du. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *2017 IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), volume 28, pages 2584–2593. IEEE, July 2017. 1

[27] Yifan Li, Haomiao Sun, Zhaori Liu, and Hu Han. Affective behaviour analysis using pretrained model with facial priori, 2022. 2, 3

[28] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-Supervised Representation Learning From Videos for Facial Action Unit Detection. In *CVPR*, pages 10924–10933, 2019. 1

[29] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016. 3

[30] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. 3

[31] Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen, and Hanzi Wang. Feature decomposition and reconstruction learning for effective facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7660–7669, 2021. 1

[32] Andrey V. Savchenko. Emotieffnet facial features in uni-task emotion recognition in video at abaw-5 competition, 2023. 6

[33] Ying Shu, Yan Yan, Si Chen, Jing-Hao Xue, Chunhua Shen, and Hanzi Wang. Learning Spatial-Semantic Relationship for Facial Attribute Recognition With Limited Labeled Data. In *CVPR*, pages 11916–11925, 2021. 1

[34] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020. 1

[35] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 1

[36] Fanglei Xue, Yifan Sun, and Yi Yang. Unsupervised facial expression representation learning with contrastive local warping, 2023. 1, 2, 3, 5

[37] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2412–2418, 2022. 2

[38] Fanglei Xue, Qiangchang Wang, and Guodong Guo. TransFER: Learning Relation-aware Facial Expression Representations with Transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, Mar. 2021. 1

[39] Fanglei Xue, Qiangchang Wang, Zichang Tan, Zhongsong Ma, and Guodong Guo. Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Transactions on Affective Computing*, 2022. 6

[40] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 1

[41] Tenggan Zhang, Chuanhe Liu, Xiaolong Liu, Yuchen Liu, Liyu Meng, Lei Sun, Wenqiang Jiang, Fengyuan Zhang, Jinming Zhao, and Qin Jin. Multi-task learning framework for emotion recognition in-the-wild. In *European Conference on Computer Vision*, pages 143–156. Springer, 2023. 2, 3

[42] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. *arXiv preprint arXiv:2107.03708*, 2021. 2

[43] Wei Zhang, Xianpeng Ji, Keyu Chen, Yu Ding, and Changjie Fan. Learning a Facial Expression Embedding Disentangled From Identity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6759–6768, 2021. 1

[44] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Multi-modal facial affective analysis based on masked autoencoder, 2023. 6

[45] Wei Zhang, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. *arXiv preprint arXiv:2203.12367*, 2022. 2

[46] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative Uncertainty Learning for Facial Expression Recognition. In *NeurIPS 2021*, page 12, 2021. 1

[47] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn From All: Erasing Attention Consistency for Noisy Label Facial Expression Recognition. In *ECCV*. arXiv, 2022. 1

[48] Ziyang Zhang, Liuwei An, Zhishun Cui, Ao xu, Tengteng Dong, Yueqi Jiang, Jingyi Shi, Xin Liu, Xiao Sun, and Meng Wang. Facial affect recognition based on transformer encoder and audiovisual fusion for the abaw5 challenge, 2023. 6

[49] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Continuous emotion recognition based on tcn and transformer, 2023. 6