Multi-Task Multi-Modal Self-Supervised Learning for Facial Expression Recognition

Marah Halawa, Florian Blume, Pia Bideau, Martin Maier, Rasha Abdel Rahman, Olaf Hellwich

arXiv:2404.10904v2 [cs.CV] 4 Sep 2024

Multi-Task Multi-Modal Self-Supervised Learning for Facial Expression Recognition

Marah Halawa[1,4*], Florian Blume[1,4*], Pia Bideau[2], Martin Maier[3,4]
Rasha Abdel Rahman[3,4], Olaf Hellwich[1,4]
[1]Technische Universität Berlin, [2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK,
[3]Humboldt-Universität zu Berlin, [4]Science of Intelligence, Research Cluster of Excellence, Berlin
{marah.halawa, florian.blume, olaf.hellwich}@tu-berlin.de
{pia.bideau}@inria.fr, {martin.maier, rasha.abdel.rahman}@hu-berlin.de

Abstract

Human communication is multi-modal; e.g.*, face-to-face interaction involves auditory signals (speech) and visual signals (face movements and hand gestures). Hence, it is essential to exploit multiple modalities when designing machine learning-based facial expression recognition systems. In addition, given the ever-growing quantities of video data that capture human facial expressions, such systems should utilize raw unlabeled videos without requiring expensive annotations. Therefore, in this work, we employ a multi-task multi-modal self-supervised learning method for facial expression recognition from in-the-wild video data. Our model combines three self-supervised objective functions: First, a multi-modal contrastive loss, that pulls diverse data modalities of the same video together in the representation space. Second, a multi-modal clustering loss that preserves the semantic structure of input data in the representation space. Finally, a multi-modal data reconstruction loss. We conduct a comprehensive study on this multi-modal multi-task self-supervised learning method on three facial expression recognition benchmarks. To that end, we examine the performance of learning through different combinations of self-supervised tasks on the facial expression recognition downstream task. Our model* ***ConCluGen*** *outperforms several multi-modal self-supervised and fully supervised baselines on the CMU-MOSEI dataset. Our results generally show that multi-modal self-supervision tasks offer large performance gains for challenging tasks such as facial expression recognition, while also reducing the amount of manual annotations required. We release our pre-trained models as well as source code publicly* [1].

1. Introduction

Facial expression recognition (FER) is a fundamental task for successful everyday human social interaction, and human-computer interaction [1]. Rooted in the context-sensitive and top-down manner of human perception, how we perceive an expression can change with (affective) context and prior knowledge [7, 18, 53], and other various other factors [58]. The same facial expression can be perceived differently depending on the situation and context [5, 16, 47]. A recent review from Maier et al. [39] highlights that to develop FER systems that align with human perception, we should consider contextual cues along with social knowledge. From a human perspective, context is inherently multi-modal, not just what is visually perceptible, as often previously treated in computer vision [31, 33, 59].

Over the last decade, deep learning aproaches have advanced the field of artificial intelligence by utilizing the massive amounts of data generated daily, *e.g*. on the internet. Large quantities of such data are multimodal, such as videos. Even though real-world (in the wild) video data helps train deeper machine-learning models, it also presents multiple challenges. Such data is usually imbalanced, noisy, and, most importantly, unlabeled. Therefore, research in deep learning and computer vision directs attention toward self-supervised learning algorithms, that aim to learn rich data representations without the need for manual labeling processes. Self-supervised learning is a form of unsupervised representation learning where the labels are extracted from the data itself, enabling label-efficient feature learning. Subsequently, the resulting models after the self-supervised learning phase can be used or adapted to downstream tasks, such as facial expression recognition. Given the growing amounts of video data that capture human facial expressions, self-supervision may allow for learning data representations from raw unlabeled video samples. Nevertheless, as mentioned above, FER is a challenging task that

*equal contribution
[1]https://github.com/tub-cv-group/conclugen

requires the integration of multi-modal contexts to align with human-level perception, since people their feelings across different modalities (visually, orally, and in other ways). Thus, not only is it essential to include these multiple modalities in learning algorithmss, but we also need to effectively model the interactions across these modalities to enhance FER. In this work, we introduce a multi-task multi-modal self-supervised approach for FER. Employing multiple tasks for self-supervision allows for learning more informative data representations, as each task enhances a certain property in the learned features, and integrating these tasks allows for capturing these complementary properties in the resulting embedding space.

Contributions. The main contributions in this paper are as follows:

- To the best of our knowledge we are the first to employ multi-task multi-modal self-supervised methods for FER.
- Our multi-modal multi-task self-supervised model *ConCluGen* (see Fig. 1) outperforms multiple self-supervised and fully supervised models on FER.
- We provide a comprehensive study on multiple self-supervised learning methods for FER.
- We make all self-supervised models (multi-modal and uni-modal) presented in this study publicly available to the research community as baselines for future studies.

2. Related work

2.1. Self-Supervised Learning

Over the last years, self-supervised learning methods advanced the research in different fields [14, 29, 43, 49], especially in computer vision applications where the amount of unlabeled data is increased rapidly due to social media, *e.g*. YouTube videos, TikTok, and TV-series. The reason is that with self-supervised learning methods, we can learn robust feature representation for the input data without the need for annotations. Such approaches generate the labels for the pretext task automatically from the data itself [32]. Different types of self-supervised methods have been developed: Some are predictive (also called generative) [20], *i.e*. anticipate and/or generate some parts of the data from another part of the data. The others are contrastive, which aim to anticipate the relations between data samples [10]. Some utilize one modality [41], while others draw on multiple modalities [54]. In this paper, we examine different self-supervised approaches and show that especially for complex computer vision applications, such as FER, models can benefit from utilizing multi-modal data in the self-supervised learning techniques.

Instance-Level Contrastive Learning. Instance-level contrastive learning has been used in multiple domains to improve learned representations from unlabeled data [10, 40]. The main goal of such algorithms is to pull representations of similar objects together and push dissimilar objects far away from each other. Despite recent achievements of instance-level contrastive methods in multiple domains, they still suffer from downsides such as class collision [60]. In downstream classification tasks, instance-level contrastive learning methods lose semantic similarity between images from the same class. To mitigate this effect, some research used graph-based methods with weak supervision to pull similar instances closer to each other [60]. Others used domain knowledge to create semantically aware positive and negative sets for contrastive learning [22]. Moreover, some research focuses on introducing special augmentation techniques that can directly improve a specific downstream task performance rather than focusing on the generality of the self-supervised pre-trained model [51]. Additionally, Doersch and Zisserman [15] study multi-task self-supervised learning for visual representation learning. They jointly learn four simple but diverse self-supervised methods (motion segmentation, colorization, Exemplar task, and relative position). In this paper, we study different self-supervised methods for FER. We also study multi-task self-supervised learning, but in different combinations compared to [15]. We also focus more on multi-modality in multi-task self-supervised learning.

Multi-Modal Self-Supervised Learning. Self-supervised learning techniques have also advanced multi-modal learning. In such an environment, the model can draw on different data modalities to learn robust representation. Some research applies contrastive techniques in a multi-modal setting for multiple modalities together [56], or from two modalities as in CLIP [46] that was first employed for text and audio representation learning, or to learn from visual and audio [42]. Moreover, Taleb et al. [55] applied a multi-modal self-supervised approach on genetic modalities, and Franceschini et al. [19] utilize facial landmarks as an additional modality for FER in videos. Additionally, a body of research work [2, 4, 9] supports multi-modal contrastive learning with multi-modal clustering to ensure that semantically similar modalities are represented close to each other in the embedding space. In this paper, we follow the same direction in our ConCluGen and ConClu models (Sec. 3.6). To the best of our knowledge, we are the first to apply such a technique for FER.

2.2. Facial Expression Recognition

FER is a challenging task in computer vision for multiple reasons, such as overlapping with other facial features such as identity [34], and overlapping between multiple expression labels [6]. Some research focuses on disentangling expression features from other facial features in the input images [21, 35]. Other research focuses on learning a multi-modal representation for the FER. In [52], Siriwardhana et al. introduced a multi-modal deep learning model for FER with a mechanism for the late fusion of multi-modal SSL features. The proposed technique is based on transformers and attention mechanisms. Our model differs from [52] in that we learn a joint representation space for all input modalities. In [19], the authors used a pairwise contrastive objective function inspired by CLIP [46] to learn unsupervised multi-modal representations from video data paired with text, audio, and facial landmarks. Our model is different from [19] in that we preserve the semantic similarity between all modalities by learning a multi-modal clustering objective that regularizes the pairwise multi-modal contrastive objective.

Other works focus on the late fusion techniques to fuse different representations from different input modalities efficiently into one multi-modal representation that is free of redundant information. The latter is less prone to overfitting in the prediction task [38]. To the best of our knowledge, we are the first to study multi-task multi-modal self-supervised learning for FER. In this work, we study the combination of multiple self-supervised losses, including multi-modal contrastive learning with clustering, instance-based contrastive learning, and generative self-supervised learning.

3. Methodology

In this section, we will explain the self-supervised pretraining techniques that we adopt in this work. Then we will explain our multi-task multimodal self-supervised model.

3.1. Problem Formulation

Each input is an utterance that consists of a segment of a video as a sequence of frames $V = (\mathbf{v}_1, \ldots, \mathbf{v_n})$, a corresponding audio spectrum $A = (\mathbf{a}_1, \ldots, \mathbf{a_m})$, and a corresponding text subtitle $T = (\mathbf{t}_1, \ldots, \mathbf{t_k})$. The task is first to learn a rich representation for the input in an unsupervised manner, then predict the facial expression class of the corresponding input video. Due to the unsupervised nature of our approach, we can build a FER model without the need for a huge dataset. We evaluate the following self-supervised methods on both multi-label and single-label FER using three datasets. Note that that both pretext and downstream tasks are independent. Each input instance I is represented by a triple input $(\mathbf{v_i}, \mathbf{t_i}, \mathbf{a_i})$ where $i \in \{0, N\}$ and N is the total number of input samples. We learn a representation for each modality by mapping each input modality to a lower dimensional space using three separate encoders as follows: A video encoder $F_v(\mathbf{v}) \Rightarrow \mathbf{D_v}$ that maps each video frame sequence to lower dimensional $\mathbf{D_v}$. A text encoder $F_t(\mathbf{t}) \Rightarrow \mathbf{D_t}$ that maps each text subtitle sequence to lower dimensional $\mathbf{D_t}$. Finally, an audio encoder $F_a(\mathbf{a}) \Rightarrow \mathbf{D_a}$ that maps each audio spectrum sequence to lower dimensional $\mathbf{D_a}$. To learn contrastive representations, we map $\mathbf{D_v}$, $\mathbf{D_t}$, and $\mathbf{D_a}$ representations into three separate projection heads (one for each modality), each consisting of two linear layers, as follows: $J_\mathbf{v}(\mathbf{D_v}) \Rightarrow \mathbf{P_v}$, $J_\mathbf{t}(\mathbf{D_t}) \Rightarrow \mathbf{P_t}$, and $J_\mathbf{a}(\mathbf{D_a}) \Rightarrow \mathbf{P_a}$. After that, to learn the multi-modal clustering, we map the output of the projection heads to one layer clustering head G (see Fig. 1), where $G(\mathbf{D_v}) \Rightarrow \mathbf{g_v}$, $G(\mathbf{D_t}) \Rightarrow \mathbf{g_t}$, and $G(\mathbf{D_a}) \Rightarrow \mathbf{g_a}$. Fig. 1 depicts an overview of this architecture. We named the encoders F *Representation Head* here.

Note that the feature extraction networks in Fig. 1 are fully fixed since we focus on improving the representations obtained from F. This means that in the previous paragraphs, $(\mathbf{v_i}, \mathbf{t_i}, \mathbf{a_i})$ are not the raw data, but the fixed output features extracted from 2D and 3D ResNet ([24] and [3]), DistilBERT [48], and DAVENet [23] respectively.

3.2. Instance-Level (Visual Only) Contrastive Learning

Instance-level contrastive learning methods are self-supervised methods that learn a representation of unlabeled data by discriminating between pairs of instances. They represent similar instances closer to each other in the embedding space, and dissimilar instances far apart. The quality of instance-level contrastive learning methods is influenced by the quality of data augmentations. This is the case because positive pairs for each instance are the augmented version of that instance, and the negative pairs are other instances in the training batch. To implement an instance-level contrastive learning method for video data, we follow the spatial augmentation process in the work of Qian et al. [45] that is consistent along the temporal dimension. Thus we will preserve the motion signal across frames, by generating random spatial augmentations across videos not across frames. We applied the following augmentations: random cropping with resizing, random horizontal flipping, color jitter, random grayscale, and Gaussian blur. Finally, we apply InfoNCE loss given as follows:

$$\ell_{\text{NCE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(sim(\mathbf{P}_\mathbf{v}^i, \mathbf{P'}_\mathbf{v}^i)/\tau)}{\sum_{k=1}^{2K} 1_{[k\neq i]} \exp(sim(\mathbf{P}_\mathbf{v}^i, \mathbf{P}_\mathbf{v}^k)/\tau)} \tag{1}$$

where τ is the positive temperature parameter, $\mathbf{P_v}$ is the representation of video frames, and sim is a similarity function. By minimizing ℓ_{NCE} we are minimizing the distance between the instance $\mathbf{P_v}$, and its augmented versions $\mathbf{P_v}'$, and maximize the distance between other instances.

3.3. Multi-Modal Contrastive Learning

To train a network that can project all modality inputs to the same embedding space, we learn a Masked Margin Softmax (MMS) loss [26] between each input modality. Masked Margin Softmax (MMS) loss maximizes the similarity of correctly paired modalities and minimizes the similarity of

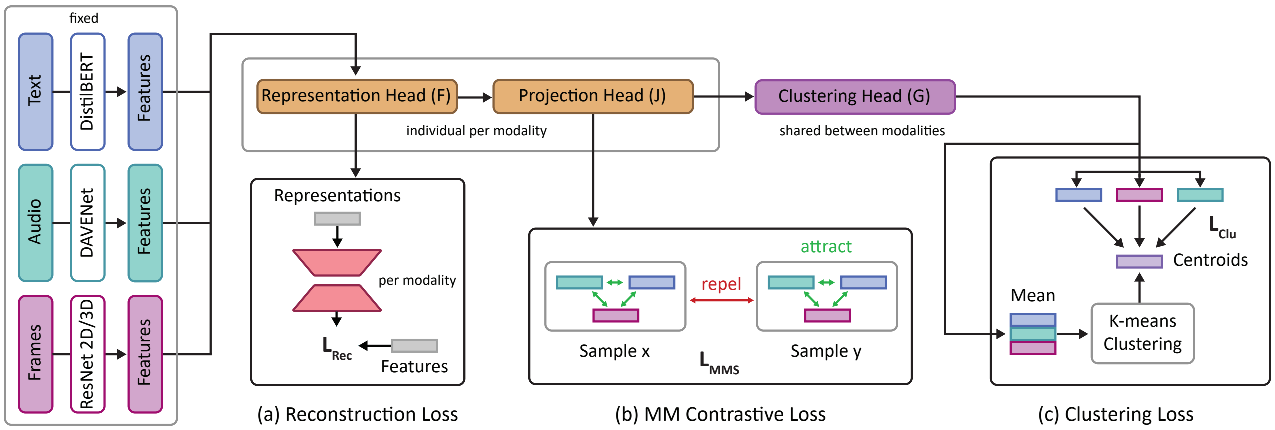


Figure 1. Overview of the architecture of the multi-task multi-modal self-supervised method. The backbone feature extractors process the input modalities in blocks, which we average over the time domain. The resulting temporal features are stored on disk, *i.e.* the backbones are fully fixed. The 3-layer representation head produces the representations we want to use in downstream training. The three losses are: (a) A reconstruction loss which reconstructs the features of each modality individually. (b) The multi-modal contrastive loss encourages the representations from the projection head for modalities belonging to the same input to be represented closer to each other. (c) The multi-modal clustering loss which drives the modalities of a sample towards the centroids computed by k-means clustering. The latter uses the mean of the modalities to compute these centroids.

incorrectly paired modalities. In the case of two modalities, text, and video frames. We calculate MMS loss as follows:

$$L_{MMS_{vw}} = L_{wv} + L_{vw} \tag{2}$$

$$L_{vw} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\mathrm{e}^{\mathbf{P_{v_i}}\cdot\mathbf{P_{w_i}}-\delta}}{\mathrm{e}^{\mathbf{P_{v_i}}\cdot\mathbf{P_{w_i}}-\delta}+\sum_{\mathrm{j=1,j\neq i}}^{\mathrm{N}}\mathrm{e}^{\mathbf{P_{v_j}}\cdot\mathbf{P_{w_i}}}} \tag{3}$$

$$L_{wv} = -\frac{1}{N}\sum_{j=1}^{N}\log\frac{\mathrm{e}^{\mathbf{P_{v_j}}\cdot\mathbf{P_{w_j}}-\delta}}{\mathrm{e}^{\mathbf{P_{v_j}}\cdot\mathbf{P_{w_j}}-\delta}+\sum_{\mathrm{i=1,j\neq i}}^{\mathrm{N}}\mathrm{e}^{\mathbf{P_{v_i}}\cdot\mathbf{P_{w_j}}}} \tag{4}$$

In the case of three modalities input, we calculate the MMS loss for each modality pair, and the final loss is the sum of all MMS losses between paired modalities.

$$L_{MMS} = L_{MMS_{vh}} + L_{MMS_{vw}} + L_{MMS_{hw}} \tag{5}$$

3.4. Generative Self-Supervised Learning

In this work, we study the impact of the reconstruction objective function on the learned representation from unlabeled data. Either as the only objective function in the proxy task or as a part of multi-task self-supervised training. Since the training data is multi-modal, we follow a similar direction as Chen et al. [9] and perform the reconstruction for each modality individually. The final reconstruction loss is the sum of the reconstruction losses of all modalities. Thus we have 3 encoder-decoder models, Q_f, Q_w, and Q_h for video, text, and audio respectively. Each decoder receives a feature representation $\mathbf{D}$ from the related reconstruction head. In the following you can see the reconstruction loss for the visual modality:

$$L_{Rec_{\mathbf{v}}} = \mathrm{MSE}(v, Q_v(\mathbf{D_v})) \tag{6}$$

The reconstruction loss is the mean-squared error (MSE) between the output features from the encoder-decoder model and the input features. Such loss penalizes the network for creating outputs (in this case features) different from the input features. The final reconstruction loss over the 3 modalities is given as:

$$L_{Rec} = L_{Rec_v} + L_{Rec_h} + L_{Rec_w} \tag{7}$$

3.5. Multi-Modal Contrastive Learning with Clustering

This method learns a multi-modal representation utilizing the multi-modal contrastive objective function from Sec. 3.3 that encourages modalities of the same instance represented closer to each other in the feature space. Moreover, it preserves the similarity between different instances by learning a multi-modal clustering objective function beside the multi-modal contrastive object function. The objective function that guides the multi-modal learning process consists of two losses: The first one is the multi-modal contrastive loss Eq. (5), the second one is the multi-modal clustering loss that will be elaborated below.

Multi-modal clustering representation learning. To preserve the similarity between different modalities of similar instances, we formulate a clustering objective function that minimizes the distance between similar modalities. To create the clustering centroids we follow a similar direction of recent work [9] that creates a multi-modal centroid over joint multi-modal representations instead of creating a separate centroid for each modality. Thus, by using the clustering loss, we encourage the audio, visual, and text embeddings of similar videos (that share similar semantic information) to be represented closer to each other in the feature space.

As shown in Eq. (8), we compute the multi-modal representation $\mathbf{R_i}$ of each instance as the mean over the representations of each modality associated with that instance. As illustrated in Fig. 1, the clustering loss is performed over the output of the *Clustering Head* which is G (see Sec. 3.1). Assuming we have three modalities (video, text, and audio) as input then the multi-modal representation $\mathbf{R}$ for each instance i is calculated as follows:

$$R_i(\mathbf{g_v}^i, \mathbf{g_t}^i, \mathbf{g_a}^i) = (\mathbf{g_v}^i + \mathbf{g_t}^i + \mathbf{g_a}^i)/3 \tag{8}$$

After estimating the cluster centroids $(\mathbf{C}_1, \ldots, \mathbf{C}_k)$, where k is the number of clusters, we estimate the multi-modal centroids using K-means over the multi-modal representations $\mathbf{R}$ by minimizing the following equation:

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \left| \mathbf{R}_i^{(j)} - \mathbf{C}_j \right| \tag{9}$$

where $\mathbf{R}_i^{(j)}$ is the multimodal representation that belongs to cluster j with centroid $\mathbf{C}j$. Finally, we learn the Multi-modal clustering loss that minimizes the distance between the multi-modal representations $(\mathbf{R}_1^{(1)}, \ldots, \mathbf{R}_n^{(k)})$ and the cluster centroids $(\mathbf{C}_1, \ldots, \mathbf{C}_k)$. This loss updates all modality encoders to encourage modality embeddings to be presented closer to each other in the feature space. The multi-modal clustering loss encourages the multi-modal representations that belong to the same cluster to be closer to their centroid thus preserving the similarity between multimodal representations.

$$\ell_{\text{Clu}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\mathrm{e}^{\mathbf{R}_i^j \cdot \mathbf{C}_j - \delta}}{\sum_{k=1}^{K} \mathrm{e}^{\mathbf{R}_\mathrm{i}^\mathrm{j} \cdot \mathbf{C}_\mathrm{k}}} \tag{10}$$

3.6. Multi-Task Self-Supervised Learning

In this section, we analyze the ability to learn rich multi-modal representations in a multi-task self-supervised learning fashion. Here we will learn multiple self-supervised tasks simultaneously without supervision, only by leveraging the structure of the multi-modal data. The intuition behind multi-task self-supervised learning is to make the most of the structure of the large unlabeled datasets which leads to rich and better generalized representation learning. In this study, we investigate the following multi-modal task combination:

1. **ConCluGen**: In this model, we jointly learn three self-supervised tasks. The first is multi-modal contrastive learning. The second is online multi-modal clustering. The third is multi-modal reconstruction. The objective function that guides the learning process in ConCluGen model is as follows:

$$L_{ConCluGen} = L_{MMS} + L_{Clu} + L_{Rec} \tag{11}$$

2. **ConClu**: In this model, we jointly learn two self-supervised tasks. First is multi-modal contrastive learning. The second is online multi-modal clustering. The objective function that guides the learning process in ConClu model is as follows:

$$L_{ConClu} = L_{MMS} + L_{Clu} \tag{12}$$

3. **ConGen**: In this model, we jointly learn two self-supervised tasks. First is multi-modal contrastive learning. The second is multi-modal reconstruction. The objective function that guides the learning process in ConGen model is as follows:

$$L_{ConGen} = L_{MMS} + L_{Rec} \tag{13}$$

4. Experiments and Analysis

In this section, we evaluate the previously detailed self-supervised methods on three facial expression recognition datasets [6, 33, 44] . First, we present the datasets used to evaluate our work. Then, we clarify the evaluation metrics (See the Appendix for implementation details). Finally, we analyze and discuss our results.

4.1. Datasets

VoxCeleb2. We pretrain all methods on the large-scale face dataset VoxCeleb2 [11], which is *not* annotated for FER. VoxCeleb2 consists of 145,000 videos of celebrities and also contains audio and subtitles. Since the videos are already focused on the face, we did not perform any cropping. For downstream task, we evaluate the CAER [33], MELD [44] and CMU-MOSEI [6] datasets. We employed a Pytorch MTCNN implementation[2] to crop out the faces (see appendix for details).

CAER. [33] consists of 13,000 videos, containing audio, of the TV series Friends. We automatically inferred subtitles

[2]https://github.com/timesler/facenet-pytorch

from the audio data. Multiple speakers can be visible within one scene. The annotation labels are the 7 basic Ekman expressions [17].

MELD. [44] is a multi-modal dataset providing frames, audio, and subtitles and is also based on the TV series Friends, similarly to CAER. Its 13,000 videos feature single individuals only, cuts in a scene result in a different sample in this dataset. It provides 7 expression labels.

CMU-MOSEI. [6] contains 3,000 videos from YouTube in which people talk mostly directly into the camera. In contrast to CAER and MELD, CMU-MOSEI offers smooth multi-label annotations for 6 emotions on a Likert scale from 0 to 3. We discretize the labels by setting every value greater than 0 to 1.

Evaluation Metrics. For the downstream task we measure accuracy (Acc.), F1 score (F1), precision and recall, all weighted by the class occurences. CAER, MELD and CMU-MOSEI are heavily imbalanced which makes weighting necessary.

4.2. Evaluation Results

In this section, we start with evaluating our model ConCluGen that we pretrained on VoxCeleb2 dataset against FER benchmarks and baselines. Then we conduct a detailed study to analyze the efficiency of the features obtained from the self-supervised methods mentioned above.

4.2.1 Benchmarking Against SOTA

In this section we are comparing the multi-task multi-modal self-supervised learning model **ConCluGen** that is pretrained on VoxCeleb, to other multi-modal self-supervised benchmarks [19, 30, 52]. Results in Tab. 1 show that ConCluGen model pretrained on VoxCeleb and finetuned on CMU-MOSEI dataset is outperforming CAE-LR [30], which is a self-supervised model that is trained on CMU-MOSEI dataset. It is also on par with Franceschini et al. [19] model performance, which uses more modalities (facial landmarks) than ConCluGen model for pretraining. It is worth mentioning that both CAE-LR [30], and Franceschini et al. [19] are pretrained on CMU-MOSEI and evaluate the downstream task on this dataset as well. However, ConCluGen model in Tab. 1 is pretrained on VoxCeleb2, and only finetuned on CMU-MOSEI.

Moreover, our ConCluGen model, both with fintuning on the CMU-MOSEI dataset on Tab. 1 and without (*i.e.* simple linear evaluation), are outperforming SSE-FT [52]. SSE-FT is an independently pre-trained SSL model for vision, audio, and text modalities. They perform late fusion using an attention mechanism. Finally, they evaluate the model on CMU-MOSEI dataset without finetuning.

Finally, results in Tab. 1 show that ConCluGen model outperforms multiple fully supervised benchmarks [6, 8, 12, 25, 38] and is on the par with [12, 28] on CMU-MOSEI.

Method	Year	Mod.	Acc.
Self-Supervised Baselines			
SSE-FT [52]	2020	VAT	55.7
CAE-LR [30]	2021	VAT	61.03
Franceschini et al. [19]	2022	VATK	66.70
ConCluGen (ours)	2024	VAT	66.48
Fully Supervised Baselines			
CMIB [38]	2022	VAT	48.2
Huynh et al. [25]	2021	VAT	57.70
Graph-MFN [6]	2018	VAT	62.35
CIA [8]	2019	VAT	62.88
MESM [12]	2021	VAT	66.80
Khare et al. [28]	2021	VAT	66.90
Wen et al. [57]	2021	VAT	82.08
Shenoy and Sardana [50]	2020	VAT	82.77

Table 1. Benchmark evaluations of our model against SOTA methods for CMU-MOSEI dataset. Modalities (*Mod.*) are **V**ideo, **A**udio, **T**ext, (facial) **K**eypoints.

Method	CMU-MOSEI	CAER	MELD
Supervised baseline	57.88	34.43	55.97
ConCluGen (linear eval.)	60.0	37.5	56.6
ConCluGen (finetuned)	66.48	50.75	60.03

Table 2. Comparison of the proposed method and baselines.

4.2.2 Supervised Baselines

The supervised baseline shown in Tab. 2 consists of the same multi-model architecture as the ConCluGen model, but with a fully-supervised objective function on the target dataset. Results in Tab. 2 show that both the linear evaluation and the finetuning approach outperform the fully supervised baseline for the three datasets. That highlights the efficiency of the multi-modal self-supervised learning method.

4.2.3 Evaluation of Self-Supervised Tasks

We start with a comparison between individual self-supervised tasks, then we evaluate different permutations of combining different self-supervised tasks, all for facial expression recognition.

Instance-level contrastive learning vs. multi-modal contrastive learning.

Results in Tab. 3, Tab. 4, and Tab. 5 show that the multi-modal contrastive learning method **Multi-Cont** (Sec. 3.3) outperforms the instance level contrastive learning **Instance-Cont** (Sec. 3.2). Thus the self-supervised multi-

Method	Acc.	F1	Prec.	Recall
ConCluGen	60.4	49.6	41.7	70.3
ConClu	57.4	43.9	44.9	60.5
ConGen	53.3	39.2	45.8	58.4
Multi-Cont	56.4	44.4	43.9	63.1
Instance-Cont	41.2	43.3	35.1	77.1
Generative	46.9	49.2	38	87.8
ConCluGen/(vision)	52.4	40.4	38.2	57.3
Multi-Cont $_{\text{Wout/text}}$	42.5	27.3	35.3	53.1
Multi-Cont $_{\text{Wout/audio}}$	53.4	47.4	41.3	75.9

Table 3. Results of different combinations of self-supervised methods on CMU-MOSEI dataset. All metrics are weighted.

Method
ConCluGen
ConClu
ConGen
Multi-Cont
Instance-Cont
Generative
put the RAVDESS stuff here292 292 Optimal Pretraining Hyperparameters Based on Downstream Accuracy In the main section of this chap
Multi-Cont $_{\text{Wout/text}}$
Multi-Cont $_{\text{Wout/audio}}$

Table 4. Results of different combinations of self-supervised methods on MELD dataset. All metrics are weighted.

Method	Acc.	F1	Prec.	Recall
ConCluGen	37.5	26.1	34.8	37.5
ConClu	36.6	24	29.4	36.6
ConGen	35.8	22.5	32.3	35.8
Multi-Cont	34.6	22.3	30.7	34.6
Instance-Cont	34.6	17.8	12	34.6
Generative	37.3	23.9	45.3	37.3
ConCluGen/(vision)	36.1	21.9	32.5	36.1
Multi-Cont $_{\text{Wout/text}}$	23.2	16.8	13.9	23.2
Multi-Cont $_{\text{Wout/audio}}$	35.1	19.3	24	35.1

Table 5. Results of different combinations of self-supervised methods on CAER dataset. All metrics are weighted.

modal contrastive method can learn better representations than the instance-level contrastive method (uni-modal). The confusion matrices for the MELD dataset in Fig. 2 indicate the quality of the classification performance of Multi-Cont model Fig. 2c across different emotion classes over Instance-Cont model Fig. 2d. See the Appendix for the confusion matrices of the CAER dataset. Moreover, we conduct an experiment to evaluate which modality is more informative to be learned along with the visual modality. Results in Tab. 3, Tab. 4, and Tab. 5 consistently show that the model can benefit more from text modality.

Contrastive learning with clustering.
Results on Tab. 4, and Tab. 5 show that learning online multi-modal clustering task along with multi-modal contrastive task during the pre-training phase, such as in **ConClu** model (Sec. 3.5), improves the results on the downstream task of FER for both MELD and CAER. Tab. 3 shows that the results are comparable between both models on CMU-MOSEI dataset. The hypothesis behind such a model is that by supporting the contrastive method with distance-based clustering, we can better capture semantic structures from the data. Fig. 2b shows how the quality of the classification in ConClu model enhanced over emotion classes on MELD dataset over Multi-Cont model (Fig. 2c).

Generative self-supervised learning vs. contrastive self-supervised learning.
To answer this question we compare the generative multi-modal model **Generative** (Sec. 3.4) that reconstructs each modality representation separately, to both multi-modal contrastive learning method **Multi-Cont** and instance-level contrastive **Instance-Cont**. Results on CMU-MOSEI and CAER datasets in Tab. 3 and Tab. 5 show that the generative loss is a robust objective function as a proxy task for FER. On the other hand, results for MELD dataset in Tab. 4 show that the **Generative** model is outperforming **Instance-Cont** but not the **Multi-Cont** model.

Multi-task multi-modal self-supervised learning.
In the previous section, results showed how powerful the generative objective function is as a proxy task compared to contrastive objective functions on some datasets. On the other hand, results in Tab. 3, Tab. 4, Tab. 5, show that for all datasets, minimizing a reconstruction loss along with a multi-modal contrastive loss **ConGen** leads to models that underperform the model that we pre-trained using only the multi-modal contrastive objective function.
Finally, we evaluate the model that learns using the multi-task multi-modal self-supervised objective function **ConCluGen** Sec. 3.6. Results in Tab. 3, and Tab. 5, show that for all datasets except for MELD Tab. 4, ConCluGen model is outperforming all other models.

How informative is the individual modality representation when learning the multi-modal multi-task self-supervised model ConCluGen?
To answer this question we evaluate the results using only the visual representation from the multi-modal multi-task model **ConCluGen** to the results of **Instance-Cont** model that only utilizes the vision modality during pretraining. Tab. 5, and Tab. 3, show that **ConCluGen/(vision)** is outperforming the model that we pre-trained using the instance

(a) ConCluGen Model. (b) ConClu Model. (c) Multi-Cont Model. (d) Instance-Cont Model.

Figure 2. Confusion metrics for MELD dataset over different self-supervised models.

level contrastive method. On the other hand, results in Tab. 4 show that we need to utilize the three modalities that are learned by **ConCluGen** model, to outperform the **Instance-Cont** model.

5. Conclusion and Future Work

In this work, we employed a multi-task multi-modal self-supervised method for facial expression recognition on three different in-the-wild datasets. Our **ConCluGen** model outperforms several multi-modal self-supervised as well as supervised baseline models. We conduct an extensive experimental study to evaluate the performance of pre-trained models with multiple self-supervised tasks.

First, we assess the gains obtained by including multiple data modalities in the self-supervised algorithm, by evaluating the performance of the model pre-trained with a multi-modal contrastive method against a model pre-trained using a contrastive method with the visual modality only, following instance-based state-of-the-art self-supervised algorithms. We found that the multi-modal contrastive method learns more informative representations than the instance-based contrastive (uni-modal method), *i.e.* the resulting model from the multi-modal method outperforms the uni-modal model on the facial expression recognition task. In another related experiment, we discarded the other data modalities from the multi-modal model, and only kept the visual modality part of the model. In this experiment, we found that even only the visual part of the multi-modal still outperforms the instance-based uni-modal model which was trained on the visual modality only. These experiments illustrate the gains obtained when including other data modalities in the self-supervised pre-training stage; the resulting features contain more information about the data - facial expressions in our case. In addition, another noteworthy observation here is that the text modality seems to enrich the resulting representations with more information about facial expressions than the audio modality.

Moreover, in another set of experiments, we evaluate the performance gains when pretraining a model with a self-supervised task that combines both a multi-modal contrastive loss and a multi-modal clustering loss simultaneously. Here, we found that combining a contrastive loss with a distance-based clustering loss encourages the model to learn more semantic structure from the data. This effect is demonstrated clearly in the improved downstream FER performance. We believe that employing a clustering loss in combination with a contrastive loss mitigates the issue of class collision that contrastive learning methods alone could encounter [60], as explained earlier.

Finally, it is noteworthy that performing the self-supervised pre-training phase in a multi-task fashion captures more semantically meaningful representations than performing each task alone. In other words, for a challenging downstream task, such as facial expression recognition in the wild, the representations learned by each individual self-supervised task alone do not seem sufficient, as demonstrated by the superior performance of **ConCluGen** against all other self-supervised models.

In future work, we would like to expand the multi-task multi-modal method for more modalities such as facial landmarks. Additionally, we aim to evaluate our multi-task multi-modal **ConCluGen** on different downstream tasks, such as facial action unit detection, face detection, and sentiment analysis. The code and the pre-trained models implemented in this work are publicly available for the research community as baselines for future studies [3].

[3] https://github.com/tub-cv-group/conclugen

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135. In addition, this work has been partially supported by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003)

References

[1] F. Abdat, C. Maaoui, and A. Pruski. Human-computer interaction using emotion recognition from facial expression. In *2011 UKSim 5th European Symposium on Computer Modeling and Simulation*, pages 196–201, 2011. 1

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 2

[3] Alex Andonian. Pretorched-x. `https://github.com/alexandonian/pretorched-x`, 2024. 3, 1

[4] Yuki Markus Asano, Mandela Patrick, Christian Rupprecht 0001, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[5] Hillel Aviezer, Noga Ensenberg, and Ran R Hassin. The inherently contextualized nature of facial emotion perception. *Current Opinion in Psychology*, 17:47–54, 2017. 1

[6] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, 2018. Association for Computational Linguistics. 2, 5, 6

[7] Julia Baum, Milena Rabovsky, Sebastian Benjamin Rose, and Rasha Abdel Rahman. Clear judgments based on unclear evidence: Person evaluation is strongly influenced by untrustworthy gossip. *Emotion*, 20(2):248–260, 2020. 1

[8] Dushyant Singh Chauhan, Md. Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Conference on Empirical Methods in Natural Language Processing*, 2019. 6

[9] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogério Schmidt Feris, David F. Harwath, James R. Glass, Michael Picheny, and Shih-Fu Chang. Multimodal clustering networks for self-supervised learning from unlabeled videos. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7992–8001, 2021. 2, 4, 5

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2

[11] J. S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 5

[12] Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. Multimodal end-to-end sparse model for emotion recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5305–5316. Association for Computational Linguistics, 2021. 6

[13] 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, 2009. 1

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2

[15] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2070–2079, 2017. 2

[16] Anna Eiserbeck, Martin Maier, Julia Baum, and Rasha Abdel Rahman. Deepfake smiles matter less—the psychological and neural impact of presumed AI-generated faces. *Scientific Reports*, 13(1):16111, 2023. 1

[17] Paul Ekman. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, 19:207–283, 1971. 6

[18] Lisa Feldman Barrett and E. Kensinger. Context Is Routinely Encoded During Emotion Perception. *Psychological science*, 2010. 1

[19] R. Franceschini, E. Fini, C. Beyan, A. Conti, F. Arrigoni, and E. Ricci. Multimodal emotion recognition with modality-pairwise unsupervised contrastive loss. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2589–2596, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 2, 3, 6

[20] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2

[21] Marah Halawa, Manuel Wöllhaf, Eduardo Vellasques, Urko Sánchez Sanz, and Olaf Hellwich. Learning disentangled expression representations from facial images. *arXiv preprint arXiv:2008.07001*, 2020. 2

[22] Marah Halawa, Olaf Hellwich, and Pia Bideau. Action-based contrastive learning for trajectory prediction. In *Computer*

Vision – ECCV 2022, pages 143–159, Cham, 2022. Springer Nature Switzerland. 2

[23] David F. Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James R. Glass. Jointly discovering visual objects and spoken words from raw sensory input. *International Journal of Computer Vision*, 128:620 – 641, 2018. 3, 1

[24] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 3, 1

[25] Van Thong Huynh, Hyung-Jeong Yang, Guee-Sang Lee, and Soo-Hyung Kim. End-to-end learning for multimodal emotion recognition in video with adaptive loss. *IEEE MultiMedia*, 28(2):59–66, 2021. 6

[26] Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. Large-scale representation learning from visually grounded untranscribed speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 55–65, Hong Kong, China, 2019. Association for Computational Linguistics. 3

[27] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 1

[28] Aparna Khare, Srinivas Parthasarathy, and Shiva Sundaram. Self-supervised learning with cross-modal transformers for emotion recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 381–388, 2020. 6

[29] Vladislav Kim, Nikolaos Adaloglou, Marc Osterland, Flavio M. Morelli, Marah Halawa, Tim König, David Gnutt, and Paula A. Marin Zapata. Self-supervision advances morphological profiling by unlocking powerful image representations. *bioRxiv*, 2024. 2

[30] Panagiotis Koromilas and Theodoros Giannakopoulos. Unsupervised multimodal language representations using convolutional autoencoders. *CoRR*, abs/2110.03007, 2021. 6

[31] Ronak Kosti, Jose Alvarez, Adria Recasens, and Agata Lapedriza. Context Based Emotion Recognition using EMOTIC Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 1

[32] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 2

[33] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-Aware Emotion Recognition Networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10142–10151, 2019. 1, 5

[34] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215, 2022. 2

[35] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2080–2089, 2018. 2

[36] I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2017. 1

[37] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 1

[38] Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, pages 1–1, 2022. 3, 6

[39] Martin Maier, Florian Blume, Pia Bideau, Olaf Hellwich, and Rasha Abdel Rahman. Knowledge-augmented face perception: Prospects for the Bayesian brain-framework to align AI and human vision. *Consciousness and Cognition*, 101: 103301, 2022. 1

[40] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013. 2

[41] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision – ECCV 2016*, pages 69–84, Cham, 2016. Springer International Publishing. 2

[42] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision*, 2018. 2

[43] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *ICML*, 2019. 2

[44] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, 2019. 5, 6

[45] R. Qian, T. Meng, B. Gong, M. Yang, H. Wang, S. Belongie, and Y. Cui. Spatiotemporal contrastive video representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6960–6970, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 3

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 2, 3

[47] James A. Russell. Reading emotions from and into faces: Resurrecting a dimensional-contextual perspective. In *The Psychology of Facial Expression*, pages 295–320. Cambridge University Press, Cambridge, 1997. 1

[48] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019. 3, 1

[49] Madeline Chantry Schiappa, Yogesh Singh Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55:1 – 37, 2022. 2

[50] Aman Shenoy and Ashish Sardana. Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 19–28, Seattle, USA, 2020. Association for Computational Linguistics. 6

[51] Yuxuan Shu, Xiao Gu, Guangyao Yang, and Benny P. L. Lo. Revisiting self-supervised contrastive learning for facial expression recognition. In *British Machine Vision Conference*, 2022. 2

[52] Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara. Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access*, 8:176274–176285, 2020. 3, 6

[53] Franziska Suess, Milena Rabovsky, and Rasha Abdel Rahman. Perceiving emotions in neutral faces: Expression processing is biased by affective person knowledge. *Social Cognitive and Affective Neuroscience*, 10(4):531–536, 2015. 1

[54] Chen Sun, Austin Myers, Carl Vondrick, Kevin P. Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7463–7472, 2019. 2

[55] Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20908–20921, 2022. 2

[56] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision – ECCV 2020*, pages 776–794, Cham, 2020. Springer International Publishing. 2

[57] Huanglu Wen, Shaodi You, and Ying Fu. Cross-modal dynamic convolution for multi-modal emotion recognition. *Journal of Visual Communication and Image Representation*, 78:103178, 2021. 6

[58] Matthias J. Wieser and Tobias Brosch. Faces in Context: A Review and Systematization of Contextual Influences on Affective Face Processing. *Frontiers in Psychology*, 3, 2012. 1

[59] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting Visual Object Recognition in Context. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12982–12991, Seattle, WA, USA, 2020. IEEE. 1

[60] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10022–10031, 2021. 2, 8

Multi-Task Multi-Modal Self-Supervised Learning for Facial Expression Recognition

Supplementary Material

6. Implementation Details

We extract features from the input modalities using pretrained fixed feature extractors. The frames for the 2D ResNet-152 [24] (ImageNet [13]) are subsampled to 1fps and for the 3D ResNet-101 [3] (Kinetics [27]) to 16fps. The audio is transformed into Mel spectrograms before processing with a DAVENet [23] pretrained on affective audio. For text, we take the last hidden layer of a DistillBERT [48] that was trained for sentiment analysis. To obtain a fixed-size feature vector per modality, we process the inputs sequentially and average the resulting vectors over time. The 2D and 3D frame features are concatenated after averaging. The result is stored on disk, *i.e.* no finetuning of the backbone feature extractors happens. We use AdamW optimizer [36] for both pertaining and downstream training, in combination with cosine annealing with warm restarts [37] as learning rate scheduling. Our batch size is 4096 and the image size is 180 by 180 pixels (cropped or resized). The size of the representations is 4096. More details are given in the supplementary material. We used the following learning rates in our experiments:

Pretraining:

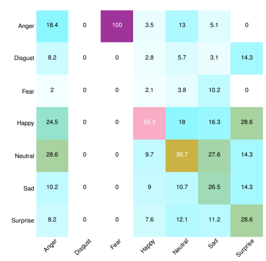
- ConCluGen: lr=0.00009, weight decay=0.00032
- ConClu: lr=0.00009, weight decay=0.00032
- ConGen: lr=0.00036, weight decay=0.00032
- Multi-Cont: lr=0.00036, weight decay=
- Instance-Cont: lr=0.00036, weight decay=00032
- Generative: lr=0.00036, weight decay=0.00032

Downstream: (We chose common hyperparameters that worked well for all methods)

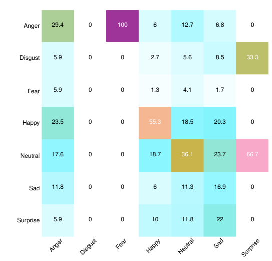
- CAER: lr=0.0061, weight decay=0.08216
- MELD: lr=0.00967, weight decay=0.00004
- MOSEI: lr=0.00996, weight decay=0.00007

For the K-means clustering, we choose a queue size of 4 (*i.e.* 4 batches were considered in the clustering), 8 clusters and started the clustering in epoch 12.

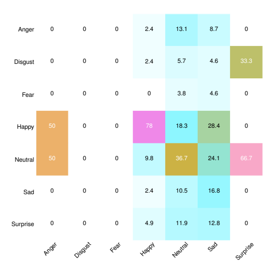
7. Confusion Matrices for CAER



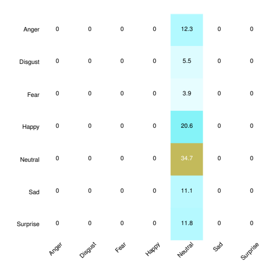
(a) ConCluGen Model.



(b) ConClu Model.



(c) Multi-Cont Model.



(d) Instance-Cont Model.

Figure 3. Confusion metrics for CAER dataset over different models.