# Cross-Attentive Adversarial Autoencoder Based on Transformer for Multi-Modal Emotion Recognition

张柏瑞（学号 202119037）
计算机科学与技术

***Abstract*—Emotion is one of the most comprehensive attributes of human beings. Instead of the uni-modal model, the multimodal model is ideal for emotion recognition as it allows for a more natural and accurate understanding of human emotions. This study proposes an enhanced multimodal adversarial autoencoder model called Cross-Attentive Adversarial Autoencoder Based on Transformer (CA-MMAAET). This model integrates cross-attention mechanisms and transformer modules to improve the fusion and representation of visual, auditory, and textual modalities. Our experimental evaluation of emotion recognition tasks shows that CA-MMAAET significantly enhances performance. The experimental results reveal that CA-MMAAET achieves an accuracy of 0.676, which significantly outperforms the baseline Multi-Modal Adversarial Autoencoder (MMAAE) with an accuracy of 0.593, as well as the Multi-Modal Adversarial Autoencoder Based on Transformer (MMAAET) and Cross-Attention Multi-Modal Adversarial Autoencoder (CA-MMAAE) with accuracies of 0.649 and 0.641, respectively. These findings demonstrate the superiority of our proposed approach in handling multimodal data, providing a more effective solution for multimodal emotion recognition tasks.**

***Keywords—Multimodal Emotion Recognition, CA-MMAAET, Cross-Attentive Mechanism, Transformer***

## I. Introduction

Multimodal emotion recognition has gained popularity in recent years due to its broad range of applications, such as human-computer interaction [1], emotional support [2], and healthcare surveillance [3]. Therefore, research on emotion recognition has attracted the focus of the research industry and the community in recent years [4].

Early research that used uni-modal approaches often failed to capture the complexity and richness of human emotions, which are inherently multimodal [5]. In contrast, the research using multimodal information to carry out emotion recognition shows a more excellent performance than the unimodal counterparts [6]. However, effectively fusing and representing multimodal data remains a significant challenge. Fortunately, recent advances in deep learning have introduced sophisticated techniques, such as attention mechanisms and transformer models, which have proven effective in various tasks, including natural language processing and computer vision. These techniques can potentially enhance multimodal emotion recognition by focusing on the most relevant features across different modalities and improving the overall representation.

This study proposes an enhanced multimodal adversarial autoencoder model called **C**ross-**A**ttention **M**ulti-**M**odal **A**dversarial **A**uto-**E**ncoder Based on **T**ransformer (CA-MMAAET). This model integrates cross-attention mechanisms and transformer modules to improve the fusion and representation of visual, auditory, and textual modalities. Our approach aims to address the limitations of existing models by leveraging the strengths of both cross-attention and transformer architectures. In summary, the contributions of this paper are mainly threefold:

- We introduce the CA-MMAAET model, which combines cross-attention mechanisms and transformer modules to improve multimodal data integration.
- We conduct extensive experiments on emotion recognition tasks to evaluate the performance of the proposed model.
- Our results demonstrate that CA-MMAAET achieves state-of-the-art performance, significantly outperforming baseline models.

## II. Related Works

There is a large volume of relevant works on multimodal emotion recognition. In this section, we only cover the most related works corresponding to the datasets and multimodal fusion models in the following.

### A. Datasets

Popular datasets for multi-modal emotion recognition or sentiment analysis include **CMU-MOSI** [7], **CMU-MOSEI** [8], **IEMOCAP** [9], **MELD** [10] [11], **CHEVAD** [12], **CHERMA** [13]. These datasets offer rich multimodal data comprising video, audio, and text, providing a robust foundation for training and evaluating emotion recognition models.

### B. Baseline Model: Multimodal Adversarial Autoencoder

The MMAAE model is a foundational approach in multimodal emotion recognition, combining adversarial training with autoencoders to extract features from various data types. It uses a multimodal autoencoder and a discriminator for generalized feature learning. However, newer techniques like Transformers and Cross-Attention, which are adept at handling complex data interactions and enhancing multimodal integration, are showing promising performance. This paper incorporates a Transformer and Cross-Attention for improved emotion recognition.

## III. Dataset Description

In this section, we give a detailed introduction to our dataset. We will present how the data is collected and annotated, the characteristics of the data, and the pre-processing of the data for model training.

Before introducing the data, we will provide definitions of some notations. Let $t$, $a$, $v$ represent the three modalities—text, audio, and vision, respectively; let $m$ denote the joint of the three modalities. They are denoted by $X_u \in \mathbb{R}^{T_u \times d_u}$ for $u \in \{t, a, v\}$ the feature sequence of the corresponding modality, where $T_u$ and $d_u$ are the sequence

length and the feature dimension. Associated with each feature sequence is its shared label $\{y_u | u \in \{m\}\}$. For our training dataset, $\{X_u^n\}_{u\in\{t,a,v\}}, \{y_u^n\}_{u\in\{m\}})$ for $n \in \{1, 2, \cdots, N\}$ is used to represent the $n$-th sample, where $N$ denotes the total number of samples.

### A. Acoustic Data: acoustic_wav2vec.pkl

The '*acoustic_wav2vec.pkl*' dataset comprises audio data processed through the wav2vec [14] framework, which is renowned for capturing robust acoustic features from raw audio waves. The length and dimensions are 128 and 512, respectively.

### B. Textual Data: textual_bert.pkl

The '*textual_bert.pkl*' dataset consists of textual data embeddings obtained from the BERT [15] model which provide a rich representation of the textual content. The length and the dimension are 36, 768 respectively.

### C. Visual Data: visual_clip.pkl

The '*visual_clip.pkl*' dataset contains visual data embeddings from the CLIP [16] model. The visual embeddings generated by CLIP capture the semantic relationships between images and their corresponding text descriptions. The length and dimension are 10 and 512, respectively.

### D. Labels: labels.pkl

The '*labels.pkl*' dataset consists of labels that categorize the data into three numerical categories: 0, 1, and 2. These labels represent negative, neutral, and positive emotions, respectively.

## IV. The Proposed Model

The proposed model, CA-MMAAET, consists of an encoder, a decoder, a discriminator, and a classifier. After the information is encoded through the encoder, the decoder, discriminator, and classifier will use it for training. Specifically, we add the **cross-attentive** mechanism in the encoder before modality fusion and the transformer block to get a rich representation of the **textual** content. In this section, we will cover four main components in the following.

**Encoder**: As shown in Figure 1, we designed an encoder with a cross-attentive mechanism and transformer block.

The transformer block was designed to gain a rich representation of the textual content. Also, we use a cross-attention block to enhance the interaction between the modalities before modality fusion to help the modality fusion in the following concatenate method.

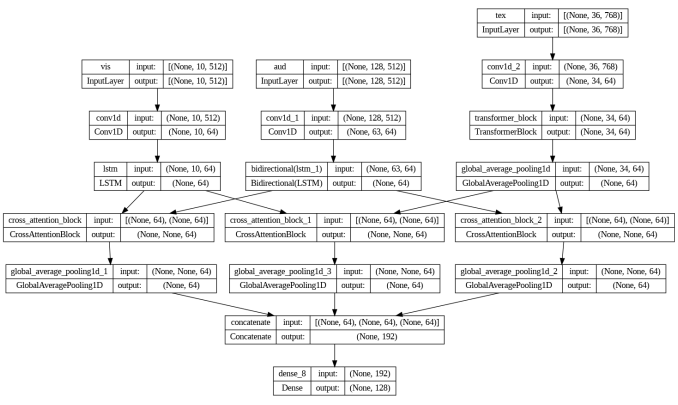
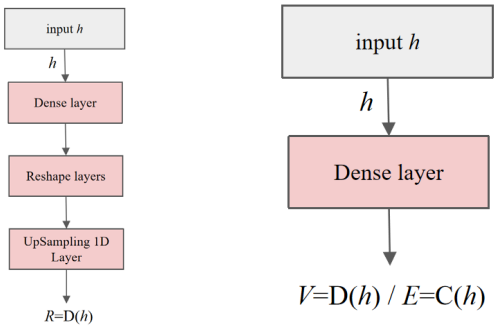



Fig. 1. The architecture of the encoder in CA-MMAAET.

Formally, given the visual features $\mathbf{X}_{\text{vis}} \in \mathbb{R}^{T_{\text{vis}} \times d_{\text{vis}}}$, audio features $\mathbf{X}_{\text{aud}} \in \mathbb{R}^{T_{\text{aud}} \times d_{\text{aud}}}$ and textual features $\mathbf{X}_{\text{tex}} \in \mathbb{R}^{T_{\text{tex}} \times d_{\text{tex}}}$. The model first processes these features through convolutional and LSTM layers to obtain intermediate representations $\mathbf{H}_{vis}, \mathbf{H}_{aud}$. $\mathbf{H}_{tex}$ on the other hand, goes through a transformer block after being processed by the convolutional layer. The cross-attention mechanism then calculates the attention weights between these modalities. For instance, between visual and audio features, the query $\mathbf{Q}$ is set to $\mathbf{H}_{vis}$, while the key $\mathbf{K}$ and value $\mathbf{V}$ are set to $\mathbf{H}_{aud}$. The attention weights are computed as $\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)$, where $d_k$ is the dimension of the key. The output of the attention mechanism is then $\text{CrossAttention}_{\text{vis}-\text{aud}} = \mathbf{AV}$. This process is extended with multi-head attention to capturing various aspects of the features by computing multiple sets of attention weights in parallel. The fused feature representations from different modalities are concatenated and passed through dense layers to form the final latent representation $\mathbf{H}_{latent}$. The cross-attention mechanism effectively leverages the complementary information across modalities, enhancing the model's capability to generate a comprehensive multimodal representation.

The transformer block processes the input tensor $\mathbf{C}_{\text{tex}} \leftarrow Convolutional\ Layer(\mathbf{H}_{tex})$ after it goes through the convolutional layer. The tensor $\mathbf{C}_{\text{tex}}$ is transformed into query, key, and value matrices. Multi-head self-attention is applied, computing attention scores and generating attention outputs for each head, concatenated and linearly transformed. The resulting tensor undergoes dropout and is combined with the original input through a residual connection followed by layer normalization, yielding $\mathbf{O}_1 = \text{LayerNorm}\left(\mathbf{C}_{\text{tex}} + \text{Dropout}(\text{MultiHeadAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}))\right)$ This is then passed through a feed-forward network consisting of two dense layers with ReLU activation, producing $\text{FFN}(\mathbf{O}_1) = \text{Dense}_2\left(\text{ReLU}\left(\text{Dense}_1(\mathbf{O}_1)\right)\right)$. The final output is obtained by another dropout, residual connection, and layer normalization: $\mathbf{O}_2 = \text{LayerNorm}\left(\mathbf{O}_1 + \text{Dropout}\left(\text{FFN}(\mathbf{O}_1)\right)\right)$. This process allows the model to effectively capture long-range dependencies and contextual information within the text, enhancing performance in downstream tasks.

**Decoder**: As shown in Figure 2(a), the decoder in the MMAAE model consists of an input layer for the latent vector, followed by two dense layers with ReLU activation that expand the dimensionality, a reshape layer to form a 3D tensor, and an upsampling layer to adjust the sequence length. This culminates in the reconstruction of the original input data. $h \rightarrow \text{Dense}(512, \text{ReLU}) \rightarrow \text{Dense}(2560, \text{ReLU}) \rightarrow \text{Reshape}(5, 512) \rightarrow \text{UpSampling1D}(2) \rightarrow \text{reconstructed output}$




*(a)* The Decoder Structure *(b)* The Discriminator & Classifier Structure

Fig. 2. The Model Structure of Decoder, Discriminator and Classifier.

**Discriminator & Classifier**: As depicted in Figure 2(b), The classifier model consists of an input layer connected to a Dense layer with 3 units and a softmax activation function: $E = \text{softmax}(\text{Dense}(h))$. Similarly, the discriminator model follows a parallel structure, with an input layer connected to a Dense layer with 1 unit and a sigmoid activation function: $V = \sigma(\text{Dense}(h))$. Here, the model can generate samples in adversarial learning.

## V. Experiment and Result

In this section, we initially compare our proposed model with conventional benchmark models to validate its effectiveness. Subsequently, we conduct ablation studies to analyze the proposed model, illustrating the distinctions between our model and its compared counterparts.

### A. Training Process

We set the epoch = 20, batch size =16. To train our model according to the process shown in Algorithm 1.

**Algorithm 1: Training Process of the Experiments**

**Input:** epoch ($e$), batch size ($b$), visual ($v$), acoustic ($a$), textual ($t$) data, label)

**Output:** The possibility of three different emotions (Negative, Neutral, and Positive)

1. Initialize the experiment model
2. Load $t$, $v$, $a$ data
3. Split data into training, validation, and test sets
4. Set hyperparameters: $e = 20$, $b = 16$
5. **for** each epoch **do**
6. Initialize empty lists for recording losses
7. **for** each batch in training set **do**
8. **Encode**: $h \leftarrow encoder(v,\ a,\ t)$
9. **Decode:** Reconstruction $R \leftarrow decoder(h)$
10. Calculate $L_{\text{rec}} = \frac{1}{N}\sum_{i=1}^{N} |R_i - \hat{R}_i|^2$
11. **Discriminate**: Validity $V \leftarrow discriminator(h)$
12. Calculate $L_{\text{adv}} = -\frac{1}{N}\sum_{i=1}^{N} V_{\text{disc}} \log\left(\hat{V}_{\text{disc}}\right) + (1 - V_{\text{disc}}) \log\left(1 - \hat{V}_{\text{disc}}\right)$
13. **Classify**: Emotions $E \leftarrow classifier(h)$
14. Calculate $L_{\text{cla}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} E_{i,k} \log\left(\hat{E}_{i,k}\right)$
15. Update the model with their respective losses
16. **end for**
17. Evaluate model performance on the validation set
18. **end for**

### B. Baseline Model: MMAAE

We trained the baseline model MMAAE according to the instructions in Algorithm 1. The loss curve was obtained, as shown in Figure 3. At the end of the training, the accuracy was **59.3%**. Specifically, we plotted the confusion matrix of the final classification result in Appendix A.

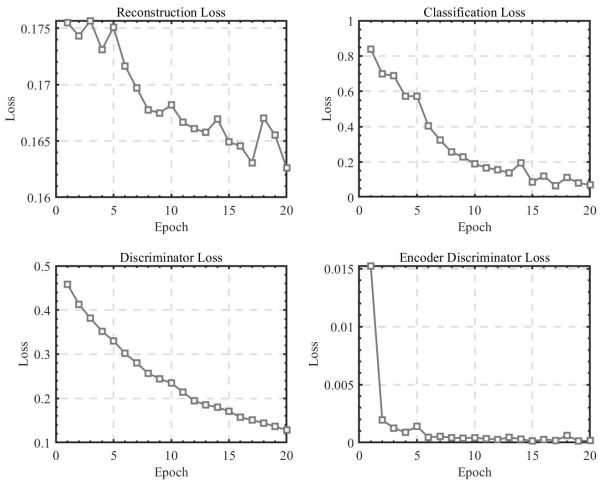



Fig. 3. The Loss Curve of MMAAE during training.

### C. Improved Model: MMAAET & CA-MMAAE

We also trained the model MMAAET, which incorporates a Transformer Block embedding in the encoder, and CA-MMAAE, which adds cross-attention between each modality before modality fusion. To evaluate the effectiveness of each modification, the training loss curves of MMAAET and CA-MMAAE are shown in Appendix B and C. The corresponding confusion matrices are listed in Appendix D and E, which show the accuracies of MMAAET and CA-MMAAE are **64.9%** and **64.1%**, respectively.

### D. The Proposed Model: CA-MMAET

Similarly, we train on our proposed model CA-MMAET, which integrates the advanced techniques in MMAAET and CA-MMAAE. Appendix F shows that we trained our model and gained the loss curve. The classification results are shown in Figure 4, with an accuracy of **67.6%**.

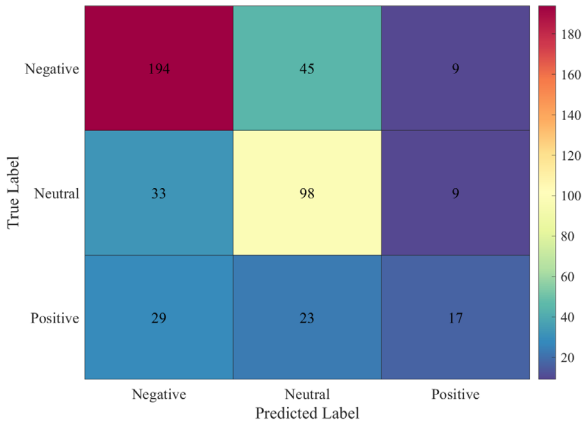



Fig. 4. The Confusion Matrix of CA-MMAET Classification Result.

### E. Performance Comparisons

In this section, we compare the baseline and our proposed model. As depicted in Figure 5, we compared the accuracy of the models. The CA-MMAAET model, which is marked with circles, performs the best. The performance of MMAAET and CA-MMAAE models is similar, but both are higher than the original model MMAAE.

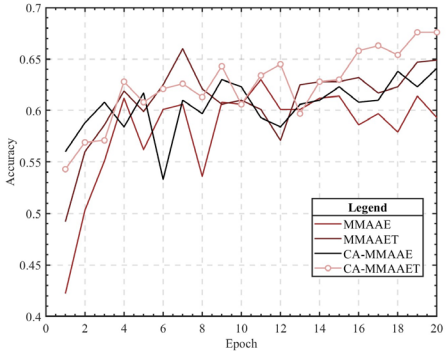



Fig. 5. The Confusion Matrix of CA-MMAAET Classification Result.

Additionally, we calculated the F1 scores for each model to evaluate their performance in sentiment classification, as shown in Table I. The MMAAE model achieved an average F1 score of 0.5022, while the CA-MMAE model scored 0.4907. The MMAET model showed improvement with an average F1 score of 0.5554, and the CA-MMAET model achieved the highest average score of 0.5790. These results highlight the effectiveness of cross-attention mechanisms and transformer-based architectures in enhancing sentiment classification performance.

TABLE I. THE F1 SCORE OF EACH MODEL

| Model | F1 Score | | | |
|---|---|---|---|---|
| | *Negative* | *Neural* | *Positive* | *Average* |
| MMAAE | 0.6985 | 0.5318 | 0.2764 | 0.5022 |
| CA-MMAE | 0.7637 | 0.5191 | 0.1895 | 0.4907 |
| MMAET | 0.7698 | 0.5844 | 0.3121 | 0.5554 |
| **CA-MMAET** | **0.7699** | **0.6400** | **0.3271** | **0.5790** |

## VI. CONCLUSIONS

We introduced the Cross-Attentive Multi-Modal Adversarial Auto-Encoder Based on Transformer (CA-MMAAET) for emotion recognition. This model leverages cross-attention mechanisms and transformer modules to integrate better and represent visual, auditory, and textual modalities. Our experiments show that CA-MMAAET significantly outperforms baseline models, achieving an accuracy of 0.676 and the highest average F1 score of 0.5790 compared to the other baselines. These results demonstrate the effectiveness of advanced cross-attention and transformer architectures in multimodal emotion recognition.

## VII. LIMITATIONS

The CA-MMAAET model has some limitations. Its performance is highly dependent on the quality and diversity of the training data. The increased computational complexity due to cross-attention mechanisms and transformer modules demands significant computational resources, which may limit real-time applications. Additionally, while effective on our datasets, the model's generalizability to other datasets and domains needs further validation. Future work should focus on optimizing efficiency and testing across a broader range of datasets.


## ACKNOWLEDGMENT

We would like to express our gratitude to Dr. Fu for his insightful lectures and valuable guidance during the classes, which let us step into the realm of Deep Learning!


## REFERENCES


[1] Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2023). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, *35*(32), 23311-23328.

[2] Tu, Q., Li, Y., Cui, J., Wang, B., Wen, J. R., & Yan, R. (2022). Misc: a mixed strategy-aware model integrating comet for emotional support conversation. *arXiv preprint arXiv:2203.13560*.

[3] Dhuheir, M., Albaseer, A., Baccour, E., Erbad, A., Abdallah, M., & Hamdi, M. (2021, June). Emotion recognition for healthcare surveillance systems using neural networks: A survey. In *2021 International Wireless Communications and Mobile Computing (IWCMC)* (pp. 681-687). IEEE.

[4] Hu, D., Wei, L., & Huai, X. (2021). Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. arXiv preprint arXiv:2106.01978.

[5] Alvarez-Gonzalez, N., Kaltenbrunner, A., & Gómez, V. (2021). Uncovering the limits of text-based emotion detection. arXiv preprint arXiv:2109.01900.

[6] Zhao, J., Zhang, T., Hu, J., Liu, Y., Jin, Q., Wang, X., & Li, H. (2022). M3ED: Multi-modal multi-scene multi-label emotional dialogue database. arXiv preprint arXiv:2205.10237.

[7] Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, *31*(6), 82-88.

[8] Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L. P. (2018, April). Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

[9] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*, 335-359.

[10] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

[11] Zahiri, S. M., & Choi, J. D. (2018, June). Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

[12] Li, Y., Tao, J., Chao, L., Bao, W., & Liu, Y. (2017). CHEAVD: a Chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, *8*, 913-924.

[13] Sun, J., Han, S., Ruan, Y. P., Zhang, X., Zheng, S. K., Liu, Y., ... & Li, T. (2023, July). Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 658-670).

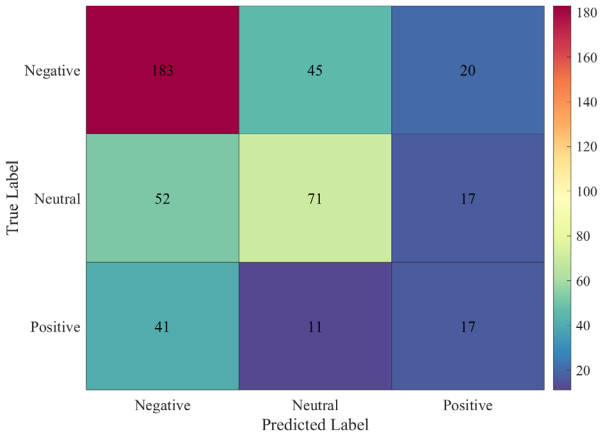
[14] Zhang, B., Lv, H., Guo, P., Shao, Q., Yang, C., Xie, L., ... & Peng, Z. (2022, May). Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6182-6186). IEEE.

[15] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

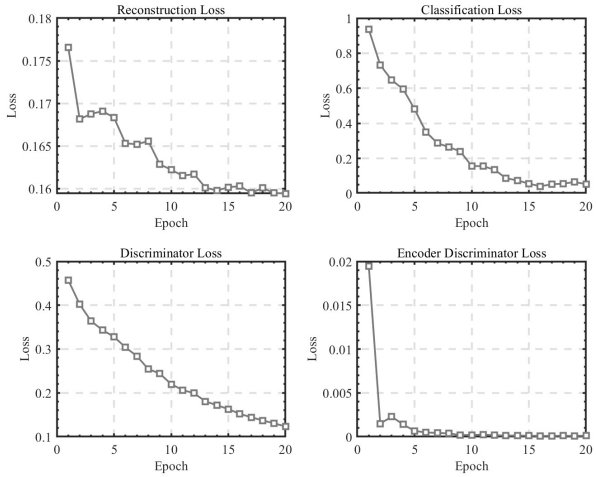
[16] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

## APPENDIX

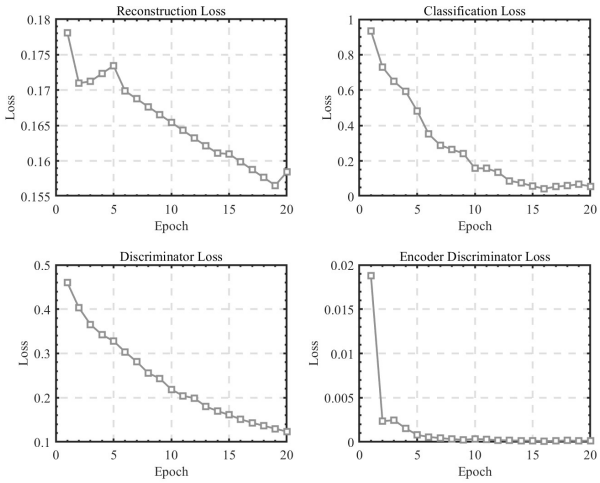
### A. The Confusion Matrix of MMAAE Classification



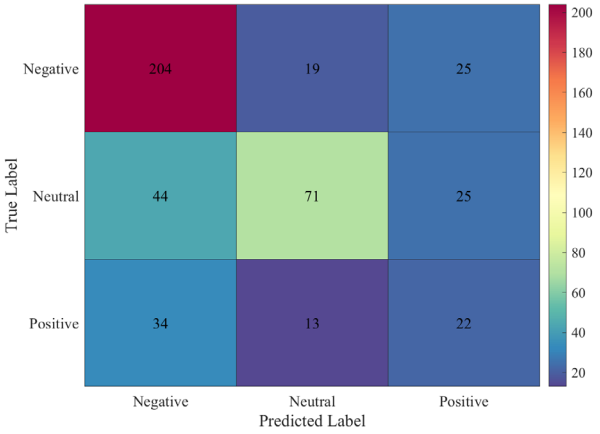

### B. The Training Loss Curve of MMAAET



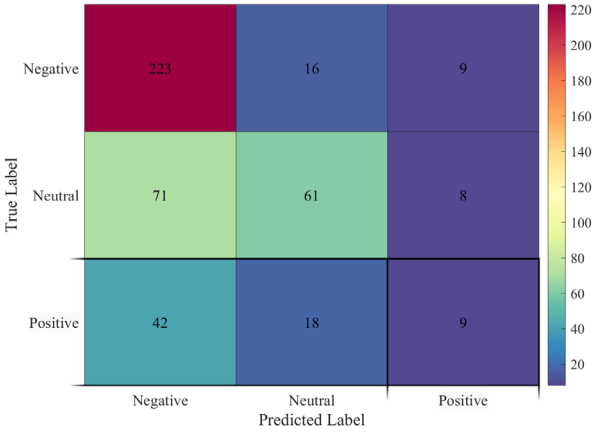

### C. The Training Loss Curve of CA-MMAAE




### D. The Confusion Matrix of MMAAET




### E. The Confusion Matrix of CA-MMAAE




### F. The Training Loss Curve of CA-MMAAET