MART: Masked Affective RepresenTation Learning via Masked Temporal Distribution Distillation

https://zzcheng.top/MART

Zhicheng Zhang[1,2] Pancheng Zhao[1,2] Eunil Park[3] Jufeng Yang[1,2,‡]
[1]VCIP & TMCC & DISSec, College of Computer Science, Nankai University
[2]Nankai International Advanced Research Institute (SHENZHEN· FUTIAN)
[3]College of Computing, Sungkyunkwan University
gloryzzc6@sina.com,pc.zhao99@gmail.com,eunilpark@skku.edu,yangjufeng@nankai.edu.cn

Abstract

Limited training data is a long-standing problem for video emotion analysis (VEA). Existing works leverage the power of large-scale image datasets for transferring while failing to extract the temporal correlation of affective cues in the video. Inspired by psychology research and empirical theory, we verify that the degree of emotion may vary in different segments of the video, thus introducing the sentiment complementary and emotion intrinsic among temporal segments. We propose an MAE-style method for learning robust affective representation of videos via masking, termed MART. First, we extract the affective cues of the lexicon and verify the extracted one by computing its matching score with video content, in terms of sentiment and emotion scores alongside the temporal dimension. Then, with the verified cues, we propose masked affective modeling to recover temporal emotion distribution. We present temporal affective complementary learning that pulls the complementary part and pushes the intrinsic one of masked multimodal features, where the constraint is set with cross-modal attention among features to mask the video and recover the degree of emotion among segments. Extensive experiments on five benchmarks show the superiority of our method in video sentiment analysis, video emotion recognition, multimodal sentiment analysis, and multimodal emotion recognition.

1. Introduction

Video emotion analysis (VEA) aims to uncover the underlying attitudes of viewers to videos [75]. With the population of video in social media [30, 80, 85], VEA has been developed for various applications in mental health protection (video violence detection [37]), online education toolkits (automatically analyzing student's attitude [4]), commercial analytics (consumer sentiment analysis [28, 32]), *etc*. Entering the deep learning era [21, 67, 72], more and more researchers pay attention to building machines for understanding human emotions [44, 47].

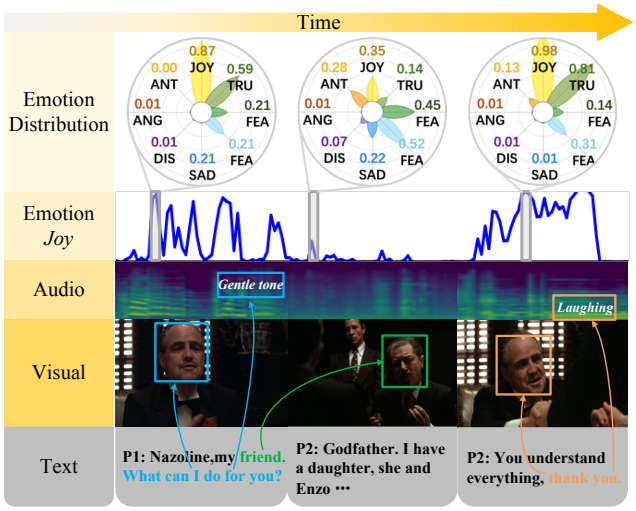


Figure 1. **A segment from *The Godfather (1972)* and the corresponding emotion scores predicted by [84].** The colored texts indicate the emotional words given by looking up the affective lexicon [7] and the rectangles show the related video content towards emotional words.

State-of-the-art VEA methods leverage the power of deep networks to extract affective representation and training networks in a fully supervised manner [48, 84]. Such a paradigm has shown its effectiveness but fatally required large-scale labeled datasets. It is time-consuming and extremely expensive to annotate emotional labels of videos [81]. Especially when considering the emotional nature of subjective perception [83], a video is required to be annotated multiple times (*e.g.*, 6 in PERR [16] and 10 in

‡ Corresponding Author.

VideoEmotion8 [31]). As a compromise, existing deep networks for VEA are initialized from the one pretrained by object classification in ImageNet [11] or action recognition in Kinetics-400 [8]. However, these methods are subject to introducing extra large-scale datasets.

A natural intuition is to consider leveraging self-supervised learning in the video emotion datasets. The key to self-supervised learning is to train the model by solving the well-designed proxy task, with which the pseudo label can be generated without human supervision. Recently, masked autoencoder (MAE) has been widely used due to its promising performance on mining correlation between local patches [74]. MAE [22, 62] recovers the self-carried information of the masked video, *e.g.*, low-level representation of pixels, contours, or trajectories. However, different from the common contents of objects and actions presented in the video, emotion is highly abstract due to its dependence on human cognition [52]. Therefore, there exists an affective gap [84] between low-level information and high-level affective semantics.

In this work, we are inspired by the emotion theory of *Plutchik Wheel* [53] to extract affective cues. As pointed out by psychology scientists, the carrier may convey emotion with varying degree [53]. For instance, the plot twist in the video can greatly raise the viewer's emotional changes [50, 51], which are usually associated with cues in the video, such as emotional words [2]. Besides, the empirical theory also verifies that there exists inconsistency of emotional stimuli among multi-modal video [83]. In light of these theoretical works, we utilize emotion degree along temporal dimension to verify the points on the video emotion analysis dataset. As shown in Fig. 1, the degree of emotion conveyed is largely different from the varying time periods. To this end, we exploit the temporal correlation of affective cues from multi-modalities. 1) Temporal Sentiment Complementary (TSC): In the temporal segment of a video, each uni-modal conveys the same sentiment, *i.e.*, positive or negative [84]. 2) Temporal Emotion Intrinsic (TEI): The expressed emotions vary as different modalities conflict with each other, *e.g.*, joy and surprise [19].

We propose a masked affective modeling method to learn discriminative affective representation, termed MART. The method is comprised of emotional lexicon extraction and masked emotion recovery. First, we extract cues via handcrafted features of the affective lexicon [7] and verify them by computing their matching score with video content. We propose a hierarchical verification strategy, in terms of sentiment, emotion class, and degree of emotion, to identify the matched cues alongside the temporal dimension. Then, with the multimodal verified cues, we propose masked affective modeling to recover temporal emotion distribution. We conduct masking according to the attentive mask, which is generated under the constraint of complementary and intrinsic parts of multiple modalities. Subsequently, we conduct reconstruction in both feature- and patch-level that recover temporal emotion distribution.

The contributions of this paper are two-fold: 1) We present a novel masked affective modeling (MAM) to exploit temporal affective cues among modalities for discriminative representation, which can be integrated into existing VEA methods as a plug-in module. 2) Extensive experiments on five challenging benchmark datasets demonstrate the effectiveness of our method, covering the downstream areas of video emotion classification, video sentiment analysis, multimodal emotion classification, and multimodal sentiment analysis.

2. Related Work

2.1. Video Emotion Analysis

VEA plays an important role in most downstream applications such as representation [6, 29, 55, 68, 78], detection [36, 40, 41], segmentation [25, 26, 82], parsing [42, 64, 77, 79], and restoration [87–89]. The majority of prior research in video emotion analysis primarily focuses on two main groups: the categorical emotion model and the dimensional one. In the categorical method [46, 59, 71], emotions are represented by a fixed number of emotion categories. For instance, as per Ekman's findings [13], there exist six fundamental emotions - anger, disgust, fear, joy, sadness, and surprise - which are universally recognizable. In contrast, dimensional models [56] usually represent emotions in a dimensional way, *e.g.*, valence-arousal-dominance (VAD). In recent times, deep features have showcased their superior capability for predicting emotions in videos compared to manually crafted features [46, 49, 73, 75, 76, 84]. Zhao *et al*. [84] propose the first end-to-end method to recognize emotions in user-generated videos, which uses three attention to capture the most discriminative information.

2.2. Masked Autoencoder for Video

Masked autoencoders [3, 22] have recently seen promising progress since their wide applications on language [12], vision [62], audio [18], *etc*. It learns a meaningful representation by recovering masked tokens such as words in text or frames in video. To learn the priors of video [15, 17, 20, 39, 60, 61], MAEst [14] and VideoMAE [62] both aim to explore the power of mask modeling. Leveraging the temporal consistent masking strategy of tube masking, they achieve promising performance with an extremely high mask ratio. Comparing the two methods mentioned above directly recovers the raw data, BEVT [65] first learns spatial representation by masked image modeling and extracts temporal dynamics with masked video modeling. However, BEVT learns semantic information by introducing an extra pretrained model which is limited in compute-intensive scenar-

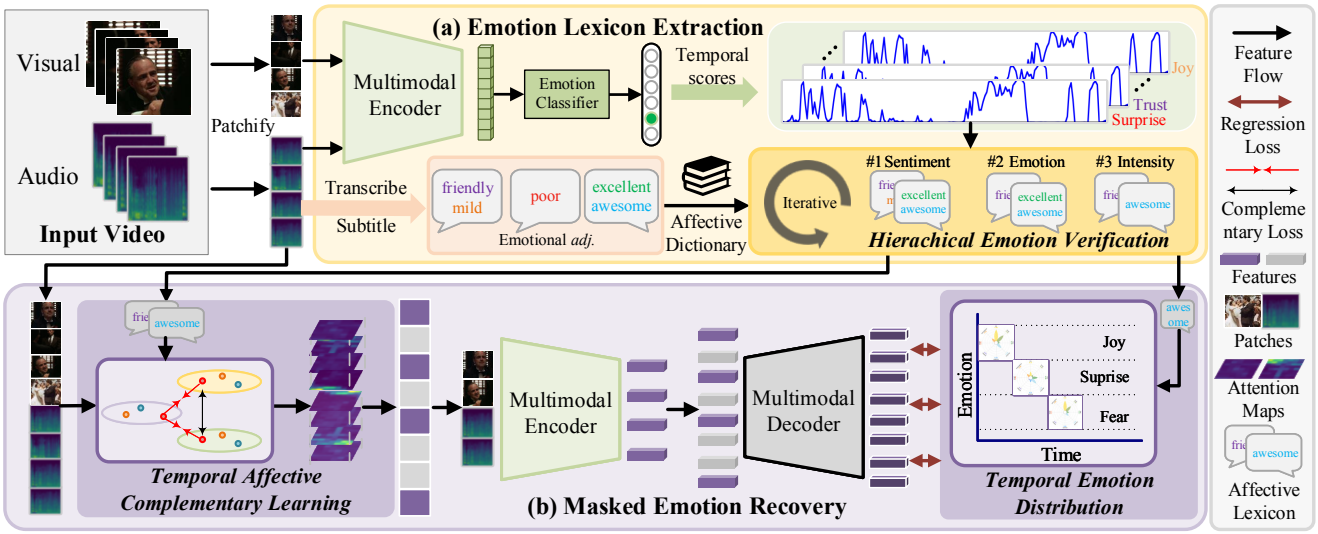


Figure 2. **The pipeline of the proposed self-supervised method MART.** Our work can be integrated into any existing SOTA VEA works by introducing our self-supervised learning method. Our method is decoupled into two modules for leveraging affective cues. In the emotion lexicon extraction module, we extract affective cues (*i.e.*, emotional words) and filter them via hierarchical lexicon verification. In the masked emotion recovery module, we first proposed to mask the part with strong emotion, which is measured by cross-attention between emotional words and video content. Then, we train the multimodal network to focus on affective segments of video and recover the corresponding temporal emotion distribution.

ios. To avoid such burden of using large pre-trained models, MaskFeat [66] introduces hand-crafted features HOG as the recovery target to learn visual knowledge.

3. Methodology

3.1. Revisiting MAE

Aiming to recover the temporal emotion distribution of video, our proposed self-supervised method is built on top of classic MAE [22]. Given an input video with T frames $v \in \mathbb{R}^{T\times d}$, where $d = HW$ denotes the size of the flattened frame. MAE split the frames into masked frames v_m and unmasked frames v_s by the binary mask $m \in \{0,1\}^T$, where the timestamp range from 1 to T. Thus, two counterparts of video v can be yielded as

$$\begin{aligned} v_m &= v[m] \in \mathbb{R}^{T_m\times d} \\ v_u &= v[1-m] \in \mathbb{R}^{T_u\times d}, \end{aligned} \tag{1}$$

where T_m and T_u represent the temporal length of masked frames and the unmasked ones, and it meets $T = T_m + T_u$.

MAE follows a commonly used encoder-decoder architecture $h = g \circ f$. A encoder f first takes as input seen frames v_u and maps them into latent embedding space D as $z_u = f(v_u) \in \mathbb{R}^{d_e}$. d_e denotes the channel of embedding space and $\circ$ is composition. To recover the masked frames, a decoder g maps these frames into the target embedding.

According to the recovery task, the target embedding carries out semantic information from the raw pixel, HoG descriptor, discrete token, *etc*. Because predicting the corresponding semantic attribute of masked frames can benefit downstream applications. Formally, the target v_t derives from the original view by $t(v)$ and the target transforms t includes mapping of identity$(\cdot)$, hog$(\cdot)$, and semantic tokenizer$(\cdot)$ and so on. Specifically, MAE aims to minimize the distance ϕ between the target semantic embedding and recovered mask embedding.

$$\mathcal{L}_{MAE}(h) = \mathbb{E}_v \mathbb{E}_{v_m, v_s|v}\, \phi(g(f(v_u)), t(v_m)). \tag{2}$$

During testing, the trained encoder is taken as the representation extractor, and task-specific header is attached above.

In this paper, we aim to mine the affective cues in the video, for generating the mask and recovering the corresponding temporal emotion distribution.

3.2. Emotional Lexicon Extraction

Psychology scientists have verified that subtitles can raise human emotions since the text can reflect the emotional state of human character and lyrics in the background music [1, 43]. In practice, multiple VEA pieces of research [10, 33, 46] also show the effectiveness of using language carried from the video due to its rich information.

• **Feature Extraction and Temporal Emotion Score** We first extract the feature of video v by any VEA methods. The visual input and audio input are mapped into joint embedding space by multimodal encoder f of VEA methods as $z \in \mathbb{R}^{T\times d_e}$. Temporal max-pooling is used for yielding global video representation z_v. Then, an emotion classifier is used to recognize the emotion from C categories as

$\overline{z_e} = \sigma(fc(z_v))$, where $fc(\cdot)$ and σ denote fully connected layer and softmax function, respectively. The classification loss from VEA methods is computed as $\mathcal{L}_{cls}$. To demonstrate the temporal emotion scores of the video, we borrow the definition of class activation map [86] to compute scores alongside temporal dimension as

$$D_{tcam} = [(\overline{z_e})_{\times T}] \cdot fc(z) \in \mathbb{R}^{T \times C}, \tag{3}$$

$[(\cdot)_{\times T}]$ is T times repeat and $\cdot$ denotes Hadamard product.

• **Emotional Lexicon** To extract affective cues, we parse the text transcribed from the video and search for the corresponding emotional word w_e via commonly-used affective lexicon [7]. Following [54], we extract the linguistic vector and emotional vector for each word. The sentiment category c_s, emotion category c_e, and the degree of each emotion i_e are obtained. Instead of using a large pre-trained model to extract semantic features, the affective information is extracted by looking up the attribute of an affective dictionary for emotional words.

• **Hierarchical Lexicon Verification** While VEA methods using language have made great progress, it is limited by their dependency on highly qualified transcribed text, which requires labor-intensive human annotation and is hard to get online for real-world situations (see details in Fig. 5). Given an emotional word w_e and the corresponding video content v, we aim to verify their matching score for usage. With the extracted temporal emotion score from class activation mapping (CAM) [86], we identify the matching score of the emotion lexicon with video content, in terms of sentiment, emotion, and degree of emotion. In the first stage, we focus on the conveyed sentiment of text c_s and video $\overline{c_s}$, *i.e.*, positive and negative, which is the primitive matching at the affective level. The matching can be computed as `E1`: $\mathbb{1}(c_s = \overline{c_s})$. In the second stage, we further investigate emotion which can be recognized by users as `E2`: $\mathbb{1}(c_s = \overline{c_s}) \cdot \mathbb{1}(c_e = \overline{c_e})$. In the final stage, we compute the match of emotion degree between text i_e and video $\overline{i_e}$, as `E3`: $\mathbb{1}(c_s = \overline{c_s}) \cdot \mathbb{1}(c_e = \overline{c_e}) \cdot \mathbb{1}(\left\|i_e - \overline{i_e}\right\| < \tau)$. τ is the threshold set for filtering.

3.3. Masked Emotion Recovery

Based on the verified emotional words, we generate the multiple attention maps by temporal affective complementary learning, where we jointly mine the affective cues among multimodalities in the video as shown in Fig. 3. With the affective features and attention obtained, we encourage the model to learn the relation between video content and the conveyed emotion, since the degree of emotion may change alongside the temporal dimension as discussed in Sec. 1. To this end, we mask the affective video and recover the corresponding temporal emotion distribution.

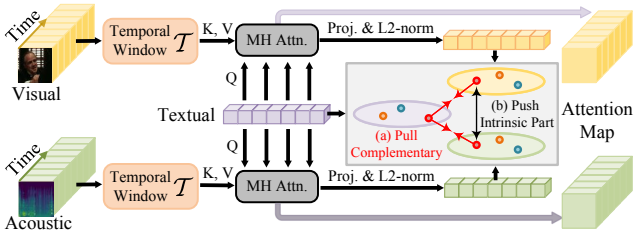


Figure 3. **Illustration of Temporal Affective Complementary Learning.** The temporal window operation splits the features according to temporal segments. Within each segment, we pull the complementary part conveying the same sentiment from multi-modalities together and push the intrinsic part expressing different emotions from cross-modalities away.

• **Temporal Affective Complementary Learning** For visual modality, we feed the linguistic feature x^l of K words into the Multi-Head Attention layer (MHA) with H heads, as the query, for retrieving the related content from visual feature $x^v \in \mathbb{R}^{(t_v h_v w_v) \times d_e}$, and yielding the lexicon-aware visual feature $x^{v \to l}$ as

$$\begin{aligned} x^{v \to l} &= \mathrm{Cat}(\mathbf{h}_1, \ldots, \mathbf{h}_H)\mathbf{W}^{v \to l} \\ \mathbf{h}_i &= a_i \mathcal{T}(x^v)\mathbf{W}_i^V \\ a_i &= \mathrm{softmax}(\frac{[x^l \mathbf{W}_i^Q][\mathcal{T}(x^v)\mathbf{W}_i^K]^\top}{\sqrt{d_k}}) \end{aligned}. \tag{4}$$

$\mathbf{W}^{v \to l}, \mathbf{W}_i^V, \mathbf{W}_i^Q, \mathbf{W}_i^K$ are learnable matrix. a_i is the attention map of the i$^{\text{th}}$ head. Cat is concatenation operation.

Symmetrically, the lexicon-aware acoustic feature $x^{a \to l}$ is derived from the acoustic feature $x^a \in \mathbb{R}^{(t_a h_a w_a) \times d_e}$ with the guidance of emotional words. Note that the temporal resolution, *i.e.*, the number of patches, is limited in attention by using the temporal window $\mathcal{T}$. We perform uniform truncation alongside temporal.

However, the matching estimated in the feature space is inaccurate, because there exists a complementary part and intrinsic part between modalities. A natural solution is to align the visual-linguistic and acoustic-linguistic to find the corresponding area in multiple modalities. We use modality-aware projection for visual and acoustic features to map them into the same semantic space, producing $x_s^{a \to l}$ and $x_s^{v \to l}$. To associate the complementary part, we push the features from video content to get close to affective-level linguistic embedding. To preserve the intrinsic part of each modality, we pull the features between visual and acoustic features. Formally, we propose the contrastive loss for emotional complementary and intrinsic learning in the k^{th} window as

$$\mathcal{L}_{tacl}^v = -\mathbb{E}_{t \in \mathcal{T}^k} \log \frac{D\left(x_s^{v \to l}[t], x^l[t]\right)}{\underbrace{D\left(x_s^{v \to l}[t], x^l[t]\right)}_{\textit{pull complement}} + \sum_{i \notin \mathcal{T}^k} \underbrace{D\left(x_s^{v \to l}[t], x_s^{a \to l}[i]\right)}_{\textit{push difference}}}. \tag{5}$$

Table 1. **Quantitative performance of the proposed method and SOTA methods for video emotion classification and video sentiment analysis.** We conduct experiments on Ekman6, VideoEmotion8, and PERR datasets. † indicates the method using captions labeled by humans. Acc^S and $F1^S$ represent the accuracy and F1 score for sentiment analysis. Acc^E and $F1^E$ represent the accuracy and F1 score for emotion analysis. SF represents supervised finetuning. SSL is self-supervised learning.

Type	Method	Year	Ekman6				VideoEmotion8				PERR			
			Acc^S	$F1^S$	Acc^E	$F1^E$	Acc^S	$F1^S$	Acc^E	$F1^E$	Acc^S	$F1^S$	Acc^E	$F1^E$
SF	Baseline [48]	NeurIPS21	68.23	68.40	48.08	46.97	72.22	72.35	43.61	42.96	60.32	57.57	56.33	52.56
SSL+SF	MoCov3 [9]	ICCV21	69.22	69.39	47.47	46.81	69.72	69.63	42.06	45.28	60.60	55.66	56.40	50.54
	AudioMAE [27]	AAAI22	69.10	69.35	45.24	44.38	68.33	68.46	41.39	40.37	54.23	52.33	51.12	48.50
	SSAST [18]	NeurIPS22	72.44	72.45	50.06	49.97	72.22	72.30	44.44	43.84	57.75	54.09	53.62	49.81
	MaskFeat [66]	CVPR22	76.14	76.15	52.41	52.40	70.00	69.57	46.39	44.76	62.29	**60.76**	57.41	54.41
	MAEst [14]	NeurIPS22	70.58	70.68	49.44	48.66	73.89	73.76	48.89	47.53	60.53	55.96	56.60	50.99
	TCL† [70]	CVPR22	69.22	69.38	46.85	46.35	70.00	69.42	44.44	41.93	**62.63**	60.18	**58.43**	54.61
	MILAN† [23]	ArXiv23	72.68	72.78	50.80	50.40	72.50	72.29	47.22	45.72	61.34	58.00	57.62	53.57
	VMAE [63]	CVPR23	70.09	70.24	50.93	50.07	**75.00**	**75.02**	49.44	46.62	61.00	56.90	56.80	51.71
	MGMAE [24]	ICCV23	76.20	76.01	51.98	52.36	69.82	70.31	49.28	45.88	61.31	61.22	55.37	56.19
	MART	CVPR24	**79.10**	**79.24**	**52.66**	**52.64**	73.61	73.58	**50.83**	**48.22**	61.88	59.25	57.89	**54.65**

Meanwhile, the loss $\mathcal{L}^a_{tacl}$ for acoustic modality can be computed symmetrically for the overall loss $\mathcal{L}_{tacl}$. We use the contrastive distance with temperature τ as $D(z, z') = \exp(z \cdot z'/\tau)$ and $\cdot$ denotes dot product.

• **Emotion-oriented Masking** To borrow the power of affective semantics, we leverage the attention map generated from the above module. With the semantic aligned feature space, we use the crossmodal attention map between language token and video content tokens, *i.e.*, visual and audio tokens of a_i from the i^{th} attention head in Equ. 4. With such, the affective attention map is computed as

$$A = \frac{1}{H}\sum_{i=1}^{H} a_i \in \mathbb{R}^{K \times d_m}, \tag{6}$$

where d_m is the resolution of the corresponding modality, *i.e.*, $h_v w_v$ for visual features and $h_a w_a$ for audio features. The attention map indicates the matching possibility of affective cues corresponding to the feature in d_m positions.

To locate the affective patch, we sort the attention map in descending order and get the corresponding index idx_i of i^{th} token. We mask high attention tokens to encourage the model to mine high-level affective cues and recover them. The maximum number of masked tokens is M. Besides, due to the abstract nature of emotion, recovering affective tokens is challenging, hiding too many tokens leads to collapse. To this end, we reveal S tokens with the highest scores that can provide cues for recovering. Further, we mask tokens in a progressive manner. That is, the fixed masking ratio is instead by an indicator that is computed by the proportion of training progress. In the early stage, we apply a low masking ratio to avoid collapse, while enlarging it as training iterations increase to encourage the model to learn robust affective representation. Finally, the masking strategy can be formulated as

$$m_i = \begin{cases} 1, S < idx_i \leq \lceil \alpha \cdot M \rceil + S, \\ 0, otherwise. \end{cases}, \tag{7}$$

where m_i decides whether the i^{th} token is to be masked and α is the proportion of the current epoch.

• **Emotion Distribution Recovery** Unlike objects in the image, we recover the temporal emotion distribution based on the temporal scores. We are inspired by the emotion theory of the Plutchik Wheel and modeling the emotion distribution for each temporal segment. We conduct both feature-level and patch-level reconstruction. The former reconstruction learns the high-level emotion distribution and the latter one improves low-level context understanding. On the one hand, we minimize the Euclidean distance between spatiotemporal model prediction and original video patches. On another hand, we apply the pooling function alongside spatial dimension and yield temporal scores. The scores are further computed with temporal distribution by KL divergence. To this end, we argue that a reconstruction task can make the model avoid overfitting on patch prediction, and improve the quality of extracted representation.

• **Discussion on $\mathcal{L}_{tacl}$** Here, we give the explanation for complementary learning from the perspective of marginal feature learning. When optimizing $\mathcal{L}_{tacl}$, the distance between emotional words and video content is minimized, while the distance between different video content is maximized to preserve the intrinsic part. For the loss of the visual part, we have the following lemma

$$\begin{aligned} \mathcal{L}^v_{tacl} &= log(1 + \exp((x_s^{v\to l} \cdot x^l - x_s^{v\to l} \cdot x_s^{a\to l})/\tau)) \\ &\approx \exp((x_s^{v\to l} \cdot x^l - x_s^{v\to l} \cdot x_s^{a\to l})/\tau) \\ &\approx 1 + \frac{1}{\tau}(x_s^{v\to l} \cdot x^l - x_s^{v\to l} \cdot x_s^{a\to l}) \\ &= 1 + \frac{1}{2\tau}(\left\|x_s^{v\to l} - x^l\right\|^2 - \left\|x_s^{v\to l} - x_s^{a\to l}\right\|^2). \end{aligned} \tag{8}$$

Table 2. **Quantitative performance of the proposed method and SOTA methods for multimodal emotion classification and multimodal sentiment analysis.** We conduct experiments on IEMOCAP and AffWild2 datasets.

Type	Methods	IEMOCAP				AffWild2			
		Acc^S	$F1^S$	Acc^E	$F1^E$	Acc^S	$F1^S$	Acc^E	$F1^E$
SF	Baseline [48]	67.02	66.09	65.79	65.71	65.54	59.05	63.50	56.74
SSL+SF	MoCov3 [9]	67.16	67.20	66.08	66.11	67.48	62.87	64.03	58.35
	MaskFeat [66]	66.37	66.04	65.50	65.19	66.30	61.38	63.87	58.36
	MAEst [14]	67.78	67.74	66.19	65.66	65.83	61.46	63.40	58.35
	SSAST [18]	63.95	64.34	62.46	61.66	65.99	59.47	63.79	55.81
	TCL† [70]	66.18	66.26	65.53	65.57	65.99	59.06	64.18	56.40
	MILAN† [23]	67.34	67.32	66.11	66.08	66.07	60.09	64.34	57.65
	VMAE [63]	65.82	65.95	65.35	65.42	66.71	61.32	65.14	58.69
	MART	**67.45**	**67.25**	**66.29**	**66.14**	**68.82**	**62.91**	**65.23**	**59.41**

Then, the loss function can be reformulated as

$$\begin{aligned}\mathcal{L}_{tacl} &= \mathcal{L}^v_{tacl} + \mathcal{L}^a_{tacl} \\ &\propto 4\tau + \left\|x_s^{a\to l} - x^l\right\|^2 + \left\|x_s^{v\to l} - x^l\right\|^2 - 2\left\|x_s^{v\to l} - x_s^{a\to l}\right\|^2. \end{aligned} \tag{9}$$

From the perspective of metric learning, we build a distance constraint for semantic feature alignment. The L2 distance between video content features, *i.e.*, $x_s^{v\to l}, x_s^{a\to l}$, is larger than the ones between video content features and emotional textual features x^l, with a margin of 4τ. To this end, $\mathcal{L}_{tacl}$ can be viewed as an extended form of triplet loss [57] for multimodal data.

4. Experiment

4.1. Dataset and Evaluation Metric

We evaluate our method on five benchmark datasets including VideoEmotion8 [31], Ekman6 [69], PERR [16], IEMOCAP [5], and AffWild2 [35]. **VideoEmotion8** dataset comprises 1,101 videos sourced from users of leading video-sharing platforms, with a minimum of 100 videos represented in each emotional category. The videos are collected from YouTube and Flicker. **Ekman6** dataset, derived from social video-sharing platforms, encompasses a robust collection of 1,637 videos, ensuring a minimum representation of 221 videos per emotional category. **PERR** dataset is built for story-like videos such as dramas and movies. To investigate emotion in social relations, each video is labeled pair-wise emotion with five human emotion categories: neutral, mild, intimate, hostile, and tense. **IEMOCAP** dataset collects conversation videos between two speakers. The conversation videos come from five sessions, which have multiple scripted plays and spontaneous dialogues. **AffWild2** dataset is an extended video datasset labeled by continuous VAD. It contains 558 videos in total with 2,786,201 frames. We measure the performance on five datasets. The overall classification accuracy and F1 score are reported for emotion, as well as the corresponding binary sentiment categories.

Table 3. **Component-wise ablation studies on Ekman6 dataset.** Msk and Ver represent emotion-oriented masking and hierarchical lexicon verification, respectively.

Module		Class						Acc^E
Msk	Ver	Ang.	Dis.	Fea.	Joy.	Sad.	Sur.	
		45.09	48.05	45.03	53.27	49.58	49.09	48.21
	✔	49.27	51.35	45.56	51.51	**58.52**	57.77	51.42
✔		**51.78**	52.55	45.57	61.46	51.56	53.30	52.16
✔	✔	51.50	**55.26**	**46.41**	**63.09**	55.86	**59.61**	**53.17**

4.2. Implementation Details

We preserve the default backbone and training loss of each baseline for performance improvement. During training, we use Adam [34] with the learning rate of 2e-4 as an optimizer. The value of weight decay is set as 0.04. For affective lexicon, we use SenticNet-5 [7] which provides a set of 100,000 natural language concepts associated with sentiment and emotion. We build the overall codebase and conduct all the experiments by PyTorch. We extract the transcripts from the given video without human correction. The codebase and transcripts are released on our homepage[1].

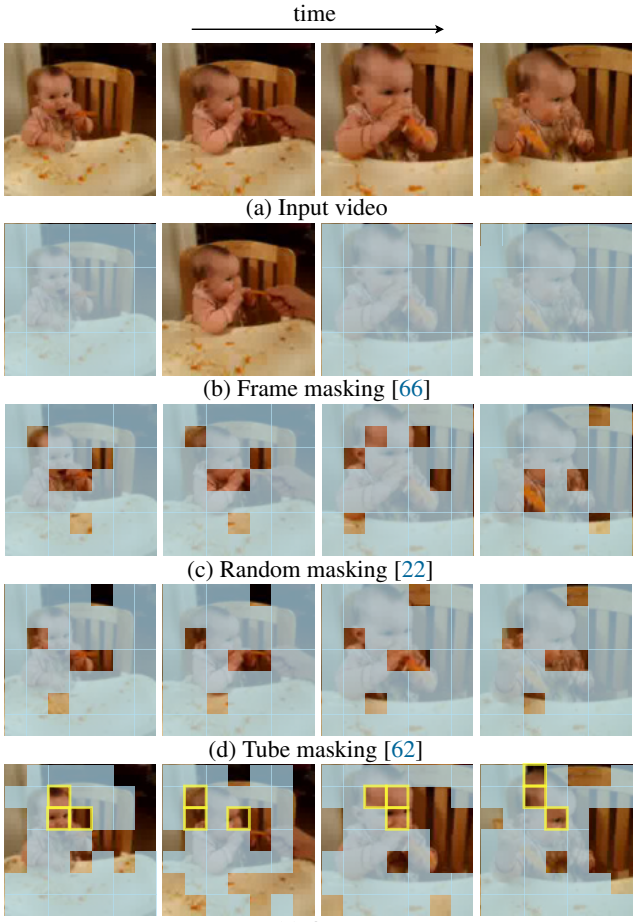
4.3. Comparison with the State-of-the-art Methods

We study the utility of affective representation by integrating our method into existing state-of-the-art VEA methods. Results are presented in Tab. 1 and Tab. 2. Each line represents the self-supervised learning method applied to the VEA baseline method. First, MART achieves competitive performance in the comparison with a margin of 3.19% to the previous best SSL method of VideoMAE in twelve metrics from two baselines. We further give detailed reasons in our ablation studies. Second, we notice a large performance improvement of 6.04% against our baseline methods. This is because MART better mines affective cues for conveyed emotion in the video.

4.4. Ablation Study

To investigate the effectiveness of emotion distribution recovery, emotion-oriented masking strategy, and hierarchical lexicon verification, we conduct component-wise ablation studies as shown in Tab. 3. We report the model performance for overall and per-class classification on the most widely used VideoEmotion8 dataset. We further discuss each component by conducting in-depth analyses of their variants to answer the following research questions. **RQ1:** What is a proper strategy for masking affective areas? **RQ2:** What is an optimal target for recovery? **RQ3:** How do mine the affective cues for boosting VEA methods?

[1] https://zzcheng.top/MART



(a) Input video

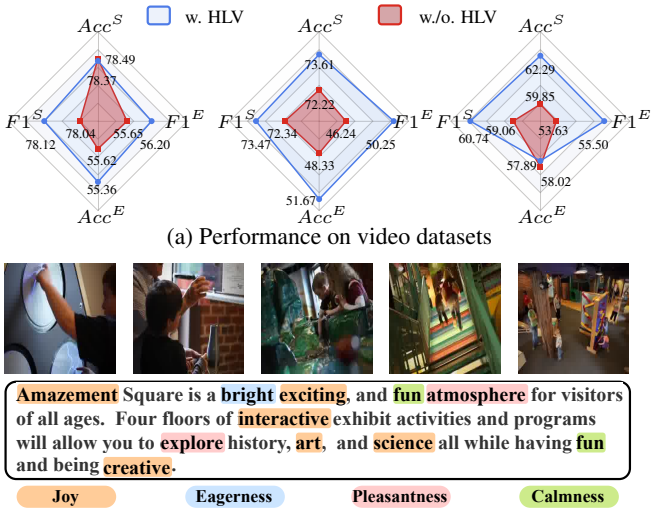
(b) Frame masking [66]

(c) Random masking [22]

(d) Tube masking [62]

(e) Ours

Figure 4. **Comparison of different masking strategies.** The blue shadowed blocks are masked to be recovered. When other masking strategies (b-d) random select the masked area, our masking strategy (e) encourages the model to focus on the affective area, guided by temporal affective complementary learning. Besides, the highly attentive areas indicated by the yellow boxes are reserved for preserving affective cues to recovery.



(a) Performance on video datasets

(b) Visualization samples

Figure 5. **Abaltion study on hierarchical lexicon verification strategy.** We report the results of with and without equipping verification strategy on video datasets in (a). The samples of verified lexicons are highlighted by colored boxes in (b).

4.5. In-depth Analysis

• **RQ1: Masking Strategy** As a key component of MAE, we explore a number of masking strategies, by incorporating MART into the SOTA VEA method [48] as shown in Tab. 4. Classic MAE adopts random masking for tubes, frames, or patches for learning occlusion-invariance representation. While its promising performance can be observed (*i.e.*, achieving a 5.92% performance boost for the baseline method), it still holds a gap between MAE and attention-guided one, due to the lack of semantic representation. From Tab. 4, we find an improvement of 3.40% between semantic-guided MAE and classic one. Meanwhile, MART learning affective representation can further improve the performance of baseline with a clear margin of 10.58%. Besides, as shown in Fig. 4, we outline different masking samples from MART and other SOTA self-supervised methods for qualitative comparison. It can be observed that our model can generate more semantic-rich maps where the face of character and trajectory of motion are better separated with less noise.

As the ratio of the masked frames increases, it is more and more difficult to recover emotion in a video and to guide model learning high-level understanding towards video. Thus, we explore studying the effect of the masking ratio for our proposed method. As shown in Tab. 4, we find that MART achieves the best performance with the progressive masking strategy. We also observe that extremely high or low masking ratios get suboptimal performance (*e.g.*, 95). We speculate the reason is double-sided. Compared to the high masking ratio, our proposed masking strategy can preserve the intrinsic structure of the affective temporal segment. In contrast, a low masking ratio reduces the difficulty of recovering masked frames and leads to a trivial solution, *i.e.*, a temporal information leakage problem [62].

• **RQ2: Recovery Target** Unlike common objects and actions that have the pre-existing vocabulary to discretize patches into tokens, emotion is required to build high-level abstract representations to mimic the procedure of human cognition. To this end, we investigate the proper feature types for providing the target affective semantics, by equipping the visual-audio-language VEA method [48] with our proposed strategy. As shown in Tab. 5, our proposed masking strategy achieves top performance by learning affective cues from multiple modalities. As the recovery target, the temporal distribution provides various and distinct supervision toward high-level cognitive emotion understanding.

Table 4. **Quantitative performance of the proposed masking strategy and others on Ekman6.**

Varient		Class						Acc^E
Masking	Ratio (%)	Ang.	Dis.	Fea.	Joy.	Sad.	Sur.	
Frame [66]	40	35.78	44.70	50.00	**58.29**	57.23	48.75	49.17
Random [14]	90	49.27	**51.35**	45.56	51.51	58.52	**57.77**	**51.42**
Tube [62]	95	**56.67**	44.70	**50.76**	44.13	**63.75**	45.51	49.07
Attention [38]	10∼50	50.28	49.73	**49.16**	57.48	55.13	57.63	49.67
Ours	10∼50	**51.50**	**55.26**	46.41	**63.09**	**55.86**	**59.61**	**53.17**

Table 5. **Quantitative performance of the proposed recovery target and others on Ekman6.**

Target	Class						Acc^E
	Ang.	Dis.	Fea.	Joy.	Sad.	Sur.	
Pixel [14]	43.06	47.78	**53.03**	50.51	50.14	49.99	48.58
HOG [66]	47.68	49.16	52.62	53.47	49.21	51.34	50.20
CLIP [23]	48.87	48.83	50.00	54.80	51.69	49.98	50.56
Ours	**51.50**	**55.26**	46.41	**63.09**	**55.86**	**59.61**	**53.17**

• **RQ3: Lexicon Verification** The quality of extracted affective lexicon is highly related to the performance of VEA methods. we conduct experiments to show the effectiveness of our proposed hierarchical lexicon verification. As shown in Fig. 5, we demonstrate the visualization samples when equipping with our verification strategy. It can be seen that the verification progress can provide consistent emotional words for MART, where the irrespective words are expelled from the extracted affective lexicon in a step-by-step manner. With the help of verified emotional words, there is a consistent performance boost in four metrics of three video emotion datasets of 2.62%.

4.6. Qualitative Analysis

Further, we present some visualization results on Ekman6 as shown in Fig. 6. As mentioned in Section 4.1, videos in this dataset are often collected from social video-sharing and can cover categories from movies to vlogs. Fig. 6 (a) demonstrate a video with its corresponding ground truth and predictions from *MBT* [48], *MBT+VideoMAE [58]*, and *MBT+MART*. Our proposed self-supervised method could recognize the conveyed emotion robustly, even with the complex change of plot and shot. In Fig. 6 (b), we further present the masked video and its corresponding recovery result by MART. The degree of conveyed emotion is recovered via understanding the video. With the occluded video, MART helps to learn the affective semantics from mining affective cues in temporal segments, such as emotional change of character and trajectory of actions. For clear contrast, we utilize the t-SNE [45] to visualize the learned affective representation in Fig. 6 (c). We observe that affective guidance shows a good extent in the feature space that MART learned. Compared to VideoMAE, the affective feature visualizations MART provided will be closer to the upper bound, resulting in enhanced discriminative of features for emotion analysis.

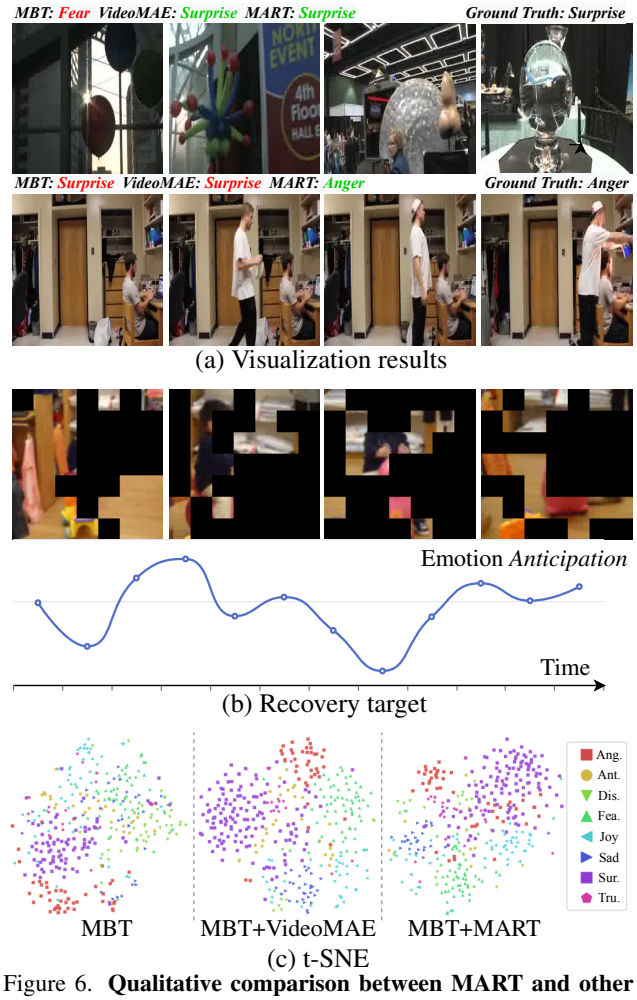


Figure 6. **Qualitative comparison between MART and other methods.** We show the prediction results (a) and the target of our emotion recovery (b), with the help of which, existing VEA methods can be improved by extracting multimodal affective cues. (c) represents the extracted affective representation of videos.

5. Conclusion

We propose an MAE-style masked affective modeling method to leverage the power of unlabeled training data. Based on the nature of temporal sentiment complementary and temporal emotion intrinsic, our method boosts existing VEA methods and outperforms SOTA methods on five challenging benchmark datasets over four downstream areas. Instead of introducing extra large-scale pretraining datasets, our self-supervised method can effectively extract affective cues from video.

6. Acknowledgments

This work was supported by Natural Science Foundation of Tianjin, China (NO.20JCJQJC00020), Fundamental Research Funds for the Central Universities, and Supercomputing Center of Nankai University (NKSC).

References

[1] Paweł Aleksandrowicz. Can subtitles for the deaf and hard-of-hearing convey the emotions of film music? a reception study. *Perspectives*, 28(1):58–72, 2020. 3

[2] Osnat Argaman. Linguistic markers and emotional intensity. *Journal of psycholinguistic research*, 39:89–99, 2010. 2

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *ICLR*, 2022. 2

[4] R Baragash and H Aldowah. Sentiment analysis in higher education: a systematic mapping review. *Journal of Physics: Conference Series*, 1860(1):012002, 2021. 1

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008. 6

[6] Qian Cai, Guo-Chong Cui, and Hai-Xian Wang. Eeg-based emotion recognition using multiple kernel learning. *MIR*, 19(5):472–484, 2022. 2

[7] Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *AAAI*, 2018. 1, 2, 4, 6

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 5, 6

[10] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *ACL*, 2020. 3

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2

[13] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, pages 45–60, 1999. 2

[14] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022. 2, 5, 6, 8

[15] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. In *arXiv*, 2022. 2

[16] Xun Gao, Yin Zhao, Jie Zhang, and Longjun Cai. Pairwise emotional relationship recognition in drama videos: Dataset and benchmark. In *ACM MM*, 2021. 1, 6

[17] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. *arXiv*, 2022. 2

[18] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. In *AAAI*, 2022. 2, 5, 6

[19] Weili Guan, Haokun Wen, Xuemeng Song, Chung-Hsing Yeh, Xiaojun Chang, and Liqiang Nie. Multimodal compatibility modeling via exploring the consistent and complementary correlations. In *ACM MM*, 2021. 2

[20] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv*, 2022. 2

[21] Olivier Habimana, Yuhua Li, Ruixuan Li, Xiwu Gu, and Ge Yu. Sentiment analysis using deep learning approaches: an overview. *SCIS*, 63(10):111192, 2020. 1

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2, 3, 7

[23] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv*, 2022. 5, 6, 8

[24] Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, and Limin Wang. Mgmae: Motion guided masking for video masked autoencoding. In *ICCV*, 2023. 5

[25] Duojun Huang, Xinyu Xiong, De-Jun Fan, Feng Gao, Xiao-Jian Wu, and Guanbin Li. Annotation-efficient polyp segmentation via active learning. *arXiv*, 2024. 2

[26] Duojun Huang, Xinyu Xiong, Jie Ma, Jichang Li, Zequn Jie, Lin Ma, and Guanbin Li. Alignsam: Aligning segment anything model to open context via reinforcement learning. In *CVPR*, 2024. 2

[27] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *NeurIPS*, 2022. 5

[28] Haeng-Jin Jang, Jaemoon Sim, Yonnim Lee, and Ohbyung Kwon. Deep sentiment analysis: Mining the causality between personality-value-attitude for analyzing business ads in social media. *Expert Systems with applications*, 40(18):7492–7503, 2013. 1

[29] Guoli Jia and Jufeng Yang. S 2-ver: Semi-supervised visual emotion recognition. In *ECCV*, 2022. 2

[30] Chenglin Jiang, Chunhong Zhang, Yang Ji, Zheng Hu, Zhiqiang Zhan, and Guanghua Yang. An affective chatbot with controlled specific emotion expression. *SCIS*, 65(10):202102, 2022. 1

[31] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. Predicting emotions in user-generated videos. In *AAAI*, 2014. 2, 6

[32] Yaping Jing, Xuequan Lu, and Shang Gao. 3d face recognition: A comprehensive survey in 2022. *CVMJ*, 9(4):657–685, 2023. 1

[33] Eesung Kim and Jong Won Shin. Dnn-based emotion recognition based on bottleneck acoustic features and lexical features. In *ICASSP*, 2019. 3

[34] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[35] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv*, 2018. 6

[36] Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *CVPR*, 2024. 2

[37] Xirong Li, Yujia Huo, Qin Jin, and Jieping Xu. Detecting violence in video using subclasses. In *ACM MM*, 2016. 1

[38] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. In *NeurIPS*, 2021. 8

[39] Zehui Lin, Ruobing Huang, Dong Ni, Jiayi Wu, and Baoming Luo. Masked video modeling with correlation-aware contrastive learning for breast cancer diagnosis in ultrasound. In *MICCAI*, 2022. 2

[40] Xin Liu, Guobao Xiao, Riqing Chen, and Jiayi Ma. Pgfnet: Preference-guided filtering network for two-view correspondence learning. *TIP*, 32:1367–1378, 2023. 2

[41] Xin Liu and Jufeng Yang. Progressive neighbor consistency mining for correspondence pruning. In *CVPR*, 2023. 2

[42] Yang Liu, Haoqin Sun, Wenbo Guan, Yuqi Xia, and Zhen Zhao. Speech emotion recognition using cascaded attention network with joint loss for discrimination of confusions. *MIR*, 20(4):595–604, 2023. 2

[43] Yun-Jhen Lu, I-Chun Kuo, and Ming-Chou Ho. The effects of emotional films and subtitle types on eye movement patterns. *Acta Psychologica*, 230:103748, 2022. 3

[44] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *CVPR*, 2021. 1

[45] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *JMLR*, 9(11):2579–2605, 2008. 8

[46] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *AAAI*, 2020. 2, 3

[47] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. Affect2mm: Affective analysis of multimedia content using emotion causality. In *CVPR*, 2021. 1

[48] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, 2021. 1, 5, 6, 7, 8

[49] Jicai Pan, Shangfei Wang, and Lin Fang. Representation learning through multimodal attention and time-sync comments for affective video content analysis. In *ACM MM*, 2022. 2

[50] Héctor J Pérez. The plot twist in tv serial narratives. *Projections*, 14(1):58–74, 2020. 2

[51] Héctor J Pérez and Rainer Reisenzein. On jon snow's death: Plot twist and global fandom in game of thrones. *Culture & Psychology*, 26(3):384–400, 2020. 2

[52] Elizabeth A Phelps. Emotion and cognition: insights from studies of the human amygdala. *Annu Rev Psychol*, 57:27–53, 2006. 2

[53] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001. 2

[54] Fan Qi, Xiaoshan Yang, and Changsheng Xu. Emotion knowledge driven video highlight detection. *TMM*, 23:3999–4013, 2020. 4

[55] Rui Qian, Weiyao Lin, John See, and Dian Li. Controllable augmentations for video representation learning. *VI*, 2(1):1–15, 2024. 2

[56] Harold Schlosberg. Three dimensions of emotion. *Psychological Review*, 61(2):81, 1954. 2

[57] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 6

[58] Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. In *ICML*, 2022. 8

[59] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. Multiple kernel learning for emotion recognition in the wild. In *ICMI*, 2013. 2

[60] Xinyu Sun, Peihao Chen, Liangwei Chen, Thomas H Li, Mingkui Tan, and Chuang Gan. M3video: Masked motion modeling for self-supervised video representation learning. *arXiv*, 2022. 2

[61] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. Vimpac: Video pre-training via masked token prediction and contrastive learning. *arXiv*, 2021. 2

[62] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 2, 7, 8

[63] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023. 5, 6

[64] Lijuan Wang, Guoli Jia, Ning Jiang, Haiying Wu, and Jufeng Yang. Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks. In *ACM MM*, 2022. 2

[65] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *CVPR*, 2022. 2

[66] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 3, 5, 6, 7, 8

[67] Changsong Wen, Guoli Jia, and Jufeng Yang. Dip: Dual incongruity perceiving network for sarcasm detection. In *CVPR*, 2023. 1

[68] Changsong Wen, Xin Zhang, Xingxu Yao, and Jufeng Yang. Ordinal label distribution learning. In *ICCV*, 2023. 2

[69] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *TAFFC*, 9(2):255–270, 2016. 6

[70] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *CVPR*, 2022. 5, 6

[71] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *CVPR*, 2018. 2

[72] Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia, and Shixia Liu. A survey of visual analytics techniques for machine learning. *CVMJ*, 7:3–36, 2021. 1

[73] Yingjie Zhai, Guoli Jia, Yu-Kun Lai, Jing Zhang, Jufeng Yang, and Dacheng Tao. Looking into gait for perceiving emotions via bilateral posture and movement graph convolutional networks. *TAC*, 2024. 2

[74] Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, Kang Zhang, and In So Kweon. A survey on masked autoencoder for self-supervised learning in vision and beyond. *arXiv*, 2022. 2

[75] Haimin Zhang and Min Xu. Recognition of emotions in user-generated videos with kernelized features. *TMM*, 20(10):2824–2835, 2018. 1, 2

[76] Jie Zhang, Yin Zhao, and Kai Qian. Enlarging the long-time dependencies via rl-based memory network in movie affective analysis. In *ACM MM*, 2022. 2

[77] Zhicheng Zhang, Song Chen, Zichuan Wang, and Jufeng Yang. Planeseg: Building a plug-in for boosting planar region segmentation. *TNNLS*, 2024. 2

[78] Zhicheng Zhang, Junyao Hu, Wentao Cheng, Danda Paudel, and Jufeng Yang. Extdm: Distribution extrapolation diffusion model for video prediction. In *CVPR*, 2024. 2

[79] Zhicheng Zhang, Shengzhe Liu, and Jufeng Yang. Multiple planar object tracking. In *ICCV*, 2023. 2

[80] Zhicheng Zhang, Lijuan Wang, and Jufeng Yang. Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In *CVPR*, 2023. 1

[81] Zhicheng Zhang and Jufeng Yang. Temporal sentiment localization: Listen and look in untrimmed videos. In *ACM MM*, 2022. 1

[82] Pancheng Zhao, Peng Xu, Pengda Qin, Deng-Ping Fan, Zhicheng Zhang, Guoli Jia, Bowen Zhou, and Jufeng Yang. Lake-red: Camouflaged images generation by latent background knowledge retrieval-augmented diffusion. In *CVPR*, 2024. 2

[83] Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. Emotion recognition from multiple modalities: Fundamentals and methodologies. *SPM*, 38(6):59–73, 2021. 1, 2

[84] Sicheng Zhao, Yunsheng Ma, Yang Gu, Jufeng Yang, Tengfei Xing, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In *AAAI*, 2020. 1, 2

[85] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer. Affective image content analysis: Two decades review and new perspectives. *TPAMI*, 44(10):6729–6751, 2022. 1

[86] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 4

[87] Shihao Zhou, Duosheng Chen, Jinshan Pan, Jinglei Shi, and Jufeng Yang. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *CVPR*, 2024. 2

[88] Shihao Zhou, Mengxi Jiang, Shanshan Cai, and Yunqi Lei. Dc-gnet: Deep mesh relation capturing graph convolution network for 3d human shape reconstruction. In *ACM MM*, 2021. 2

[89] Shihao Zhou, Mengxi Jiang, Qicong Wang, and Yunqi Lei. Towards locality similarity preserving to 3d human pose estimation. In *ACCV*, 2020. 2