How you feelin'? Learning Emotions and Mental States in Movie Scenes

Dhruv Srivastava   Aditya Kumar Singh   Makarand Tapaswi
CVIT, IIIT Hyderabad, India
https://katha-ai.github.io/projects/emotx

Abstract

Movie story analysis requires understanding characters' emotions and mental states. Towards this goal, we formulate emotion understanding as predicting a diverse and multi-label *set of emotions at the level of a movie scene and for each character. We propose EmoTx, a multimodal Transformer-based architecture that ingests videos, multiple characters, and dialog utterances to make joint predictions. By leveraging annotations from the MovieGraphs dataset [72], we aim to predict classic emotions (*e.g. *happy, angry) and other mental states (*e.g. *honest, helpful). We conduct experiments on the most frequently occurring 10 and 25 labels, and a mapping that clusters 181 labels to 26. Ablation studies and comparison against adapted state-of-the-art emotion recognition approaches shows the effectiveness of EmoTx. Analyzing EmoTx's self-attention scores reveals that expressive emotions often look at character tokens while other mental states rely on video and dialog cues.*

1. Introduction

In the movie *The Pursuit of Happyness*, we see the protagonist experience a roller-coaster of emotions from the lows of breakup and homelessness to the highs of getting selected for a coveted job. Such heightened emotions are often useful to draw the audience in through relatable events as one empathizes with the character(s). For machines to understand such a movie (broadly, story), we argue that it is paramount to track how characters' emotions and mental states evolve over time. Towards this goal, we leverage annotations from MovieGraphs [72] and train models to watch the video, read the dialog, and predict the emotions and mental states of characters in each movie scene.

Emotions are a deeply-studied topic. From ancient Rome and Cicero's 4-way classification [60], to modern brain research [33], emotions have fascinated humanity. Psychologists use of Plutchik's wheel [53] or the proposal of universality in facial expressions by Ekman [18], structure has been provided to this field through various theories. Affective emotions are also grouped into mental (affective, behavioral, and cognitive) or bodily states [13].

Figure 1. Multimodal models and multi-label emotions are necessary for understanding the story. **A**: What character emotions can we sense in this scene? Is a single label enough? **B**: Without the dialog, can we try to guess the emotions of the Sergeant and the Soldier. **C**: Is it possible to infer the emotions from the characters' facial expressions (without subtitles and visual background) only? Check the footnote below for the ground-truth emotion labels for these scenes and the supplement for an explanation of the story.

A recent work on recognizing emotions with visual context, Emotic [31] identifies 26 label clusters and proposes a *multi-label* setup wherein an image may exhibit multiple emotions (*e.g. peace, engagement*). An alternative to the categorical space, valence, arousal, and dominance are also used as three continuous dimensions [31]. Predicting a rich set of emotions requires analyzing multiple contextual modalities [31, 34, 44]. Popular directions in multimodal emotion recognition are Emotion Recognition in Conversations (ERC) that classifies the emotion for every dialog utterance [42, 54, 83]; or predicting a single valence-activity score for short ~10s movie clips [4, 45].

We operate at the level of a *movie scene*: a set of shots telling a sub-story, typically at one location, among a defined cast, and in a short time span of 30 to 60 s. Thus, scenes are considerably longer than single dialogs [54] or

Ground-truth emotions and mental states portrayed in movie scenes in Fig. 1: **A**: excited, curious, confused, annoyed, alarmed; **B**: shocked, confident; **C**: happy, excited, amused, shocked, confident, nervous.

movie clips in [4]. We predict emotions and mental states for all characters in the scene and also by accumulating labels at the scene level. Estimation on a larger time window naturally lends itself to multi-label classification as characters may portray multiple emotions simultaneously (*e.g. curious* and *confused*) or have transitions due to interactions with other characters (*e.g. worried* to *calm*).

We perform experiments with multiple label sets: Top-10 or 25 most frequently occurring emotion labels in MovieGraphs [72] or a mapping to the 26 labels in the Emotic space, created by [45]. While emotions can broadly be considered as part of mental states, for this work, we consider that *expressed emotions* are apparent by looking at the character, *e.g. surprise, sad, angry*; and *mental states* are latent and only evident through interactions or dialog, *e.g. polite, determined, confident, helpful*[1]. We posit that classification in a rich label space of emotions requires looking at multimodal context as evident from masking context in Fig. 1. To this end, we propose EmoTx that jointly models video frames, dialog utterances, and character appearance.

We summarize our contributions as follows: (i) Building on rich annotations from MovieGraphs [72], we formulate scene and per-character emotion and mental state classification as a multi-label problem. (ii) We propose a multimodal Transformer-based architecture EmoTx that predicts emotions by ingesting all information relevant to the movie scene. EmoTx is also able to capture label co-occurrence and jointly predicts all labels. (iii) We adapt several previous works on emotion recognition for this task and show that our approach outperforms them all. (iv) Through analysis of the self-attention mechanism, we show that the model learns to look at relevant modalities at the right time. Self-attention scores also shed light on our model's treatment of expressive emotions *vs.* mental states.

2. Related Work

We first present work on movie understanding and then dive into visual and multimodal emotion recognition.

Movie understanding has evolved over the last few years from person clustering and identification [6, 7, 19, 29, 46, 65] to analyzing the story. Scene detection [11, 55, 56, 58, 66], question-answering [35, 68, 77], movie captioning [57, 78] with names [50], modeling interactions and/or relationships [21, 32, 43], alignment of text and video storylines [67, 76, 84] and even long-form video understanding [75] have emerged as exciting areas. Much progress has been made through datasets such as Condensed Movies [3], MovieNet [27], VALUE benchmark (goes beyond movies) [37], and MovieGraphs [72]. Building on the annotations from MovieGraphs [72], we focus on another pillar of story understanding complementary to the above directions: identifying the emotions and mental states of each character and the overall scene in a movie.

Visual emotion recognition has relied on face-based recognition of Ekman's 6 classic emotions [18], and was popularized through datasets such as MMI [49], CK and CK+ [41, 70]. A decade ago, EmotiW [16], FER [24], and AFEW [15] emerged as challenging in-the-wild benchmarks. At the same time, approaches such as [38, 39] introduced deep learning to expression recognition achieving good performance. Breaking away from the above pattern, the Emotic dataset [31] introduced the use of 26 labels for emotion understanding in images while highlighting the importance of context. Combining face features and context using two-stream CNNs [34] or person detections with depth maps [44] were considered. Other directions in emotion recognition include estimating valence-arousal (continuous variables) from faces with limited context [69], learning representations through webly supervised data to overcome biases [48] or improving them further through a joint text-vision embedding space [73]. Different from the above, our work focuses on multi-label emotions and mental states recognition in movies exploiting multimodal context both at the scene- and character-level.

Multimodal datasets for emotion recognition have seen recent adoption. Acted Facial Expressions in the Wild [15] aims to predict emotions from faces, but does not provide any context. The Stanford Emotional Narratives Dataset [47] contains participant shared narratives of positive/negative events in their lives. While multimodal, these are quite different from edited movies and stories that are our focus. The Multimodal EmotionLines Dataset (MELD) [54] is an example of Emotion Recognition in Conversations (ERC) and attempts to estimate the emotion for every dialog utterance in TV episodes from *Friends*. Different from MELD, we operate at the time-scale of a cohesive story unit, a movie scene. Finally, closest to our work, Annotated Creative Commons Emotional DatabasE (LIRIS-ACCEDE) [4] obtains emotion annotations for short movie clips. However, the clips are quite small (8 to 12 s) and annotations are obtained in the continuous valence-arousal space. Different from the above works, we also aim to predict character-level mental states and demonstrate that video and dialog context helps for such labels.

Multimodal emotion recognition methods. RNNs have been used since early days for ERC [28, 42, 62, 74] (often with graph networks [23, 80]) as they allow effective combination of audio, visual, and textual data. Inspired by recent advances, Transformer architectures are also adopted for ERC [12, 61]. External knowledge graphs provide useful commonsense information [22] while topic modeling

[1] Admittedly it is not always easy or possible to categorize a label as an expressed emotion or a mental state, *e.g. cheerful, upset*. Using Clore *et al.* [13]'s classification, *expressed emotions* refer to affective and bodily states, while our *mental states* refer to behavioral and cognitive states.

integrated with Transformers have improved results [83]. Multi-label prediction has also been attempted by considering a sequence-to-set approach [79], however that may not scale with number of labels. While we adopt a Transformer for joint modeling, our goal to predict emotions and mental states for movie scenes and characters is different from ERC. We adapt some of the above methods and compare against them in our experiments. Close to our work, the MovieGraphs [72] emotion annotations are used to model changing emotions across the entire movie [45], and for Temporal Emotion Localization [36]. However, the former tracks one emotion in each scene, while the latter proposes a different, albeit interesting direction.

3. Method

EmoTx leverages the self-attention mechanism in Transformers [71] to predict emotions and mental states. We first define the task (Sec. 3.1) and then describe our proposed approach (Sec. 3.2), before ending this section with details regarding training and inference (Sec. 3.3).

3.1. Problem Statement

We assume that movies have been segmented automatically [55] or with a human-in-the-loop process [66, 72] into coherent *scenes* that are self-contained and describe a short part of the story. The focus of this work is on characterizing emotions within a movie scene that are often quite long (30 to 60 s) and may contain several tens of shot changes.

Consider such a movie scene $\mathcal{S}$ that consists of a set of video frames $\mathcal{V}$, characters $\mathcal{C}$, and dialog utterances $\mathcal{U}$. Let us denote the set of video frames as $\mathcal{V} = \{f_t\}_{t=1}^{T}$, where T is the number of frames after sub-sampling. Multiple characters often appear in any movie scene. We model N characters in the scene as $\mathcal{C} = \{\mathcal{P}^i\}_{i=1}^{N}$, where each character $\mathcal{P}^i = \{(f_t, b_t^i)\}$ may appear in some frame f_t of the video at the spatial bounding box b_t^i. We assume that b_t^i is empty if the character $\mathcal{P}^i$ does not appear at time t. Finally, $\mathcal{U} = \{u_j\}_{j=1}^{M}$ captures the dialog utterances in the scene. For this work, we use dialogs directly from subtitles and thus assume that they are unnamed. While dialogs may be named through subtitle-transcript alignment [19], scripts are not always available or reliable for movies.

Task formulation. Given a movie scene $\mathcal{S}$ with its video, character, and dialog utterance, we wish to predict the emotions *and* mental states (referred as labels, or simply emotions) at both the scene, $\mathbf{y}^{\mathcal{V}}$, and per-character, $\mathbf{y}^{\mathcal{P}^i}$, level. We formulate this as a multi-label classification problem with K labels, *i.e.* $\mathbf{y} = \{y_k\}_{k=1}^{K}$. Each $y_k \in \{0, 1\}$ indicates the absence or presence of the k^{th} label in the scene $y_k^{\mathcal{V}}$ or portrayed by some character $y_k^{\mathcal{P}^i}$. For datasets with character-level annotations, scene-level labels are obtained through a simple logical `OR` operation, *i.e.* $\mathbf{y}^{\mathcal{V}} = \bigoplus_{i=1}^{N} \mathbf{y}^{\mathcal{P}^i}$.

3.2. EmoTx: Our Approach

We present EmoTx, our Transformer-based method that recognizes emotions at the movie scene and per-character level. A preliminary video pre-processing and feature extraction pipeline extracts relevant representations. Then, a Transformer encoder combines information across modalities. Finally, we adopt a classification module inspired by previous work on multi-label classification with Transformers [40]. An overview of the approach is presented in Fig. 2.

Preparing multimodal representations. Recognizing complex emotions and mental states (*e.g. nervous, determined*) requires going beyond facial expressions to understand the larger context of the story. To facilitate this, we encode multimodal information through multiple lenses: (i) the video is encoded to capture where and what event is happening; (ii) we detect, track, cluster, and represent characters based on their face and/or full-body appearance; and (iii) we encode the dialog utterances as information complementary to the visual domain.

A pretrained encoder $\phi_{\mathcal{V}}$ extracts relevant visual information from a single or multiple frames as $\mathbf{f}_t = \phi_{\mathcal{V}}(\{f_t\})$. Similarly, a pretrained language model $\phi_{\mathcal{U}}$ extracts dialog utterance representations as $\mathbf{u}_j = \phi_{\mathcal{U}}(u_j)$. Characters are more involved as we need to first localize them in the appropriate frames. Given a valid bounding box b_t^i for person $\mathcal{P}^i$, we extract character features using a backbone pretrained for emotion recognition as $\mathbf{c}_t^i = \phi_{\mathcal{C}}(f_t, b_t^i)$.

Linear projection. Token representations in a Transformer often combine the core information (*e.g.* visual representation) with meta information such as the timestamp through position embeddings (*e.g.* [63]). We first bring all modalities to the same dimension with linear layers. Specifically, we project visual representation $\mathbf{f}_t \in \mathbb{R}^{D_{\mathcal{V}}}$ using $\mathbf{W}_{\mathcal{V}} \in \mathbb{R}^{D \times D_{\mathcal{V}}}$, utterance representation $\mathbf{u}_j \in \mathbb{R}^{D_{\mathcal{U}}}$ using $\mathbf{W}_{\mathcal{U}} \in \mathbb{R}^{D \times D_{\mathcal{U}}}$, and character representation $\mathbf{c}_t^i \in \mathbb{R}^{D_{\mathcal{C}}}$ using $\mathbf{W}_{\mathcal{C}} \in \mathbb{R}^{D \times D_{\mathcal{C}}}$. We omit linear layer biases for brevity.

Modality embeddings. We learn three embedding vectors $\mathbf{E}^{\mathcal{M}} \in \mathbb{R}^{D \times 3}$ to capture the three modalities corresponding to (1) video, (2) characters, and (3) dialog utterances. We also assist the model in identifying tokens coming from characters by including a special character count embedding, $\mathbf{E}^{C} \in \mathbb{R}^{D \times N}$. Note that the modality and character embeddings do not encode any specific meaning or imposed order (*e.g.* higher to lower appearance time, names in alphabetical order) - we expect the model to use this only to distinguish one modality/character from the other.

Time embeddings. The number of tokens depend on the chosen frame-rate To inform the model about relative temporal order across modalities, we adopt a discrete time binning strategy that translates real time (in seconds) to an index. Thus, video frame/segment and character box representations fed to the Transformer are associated with their

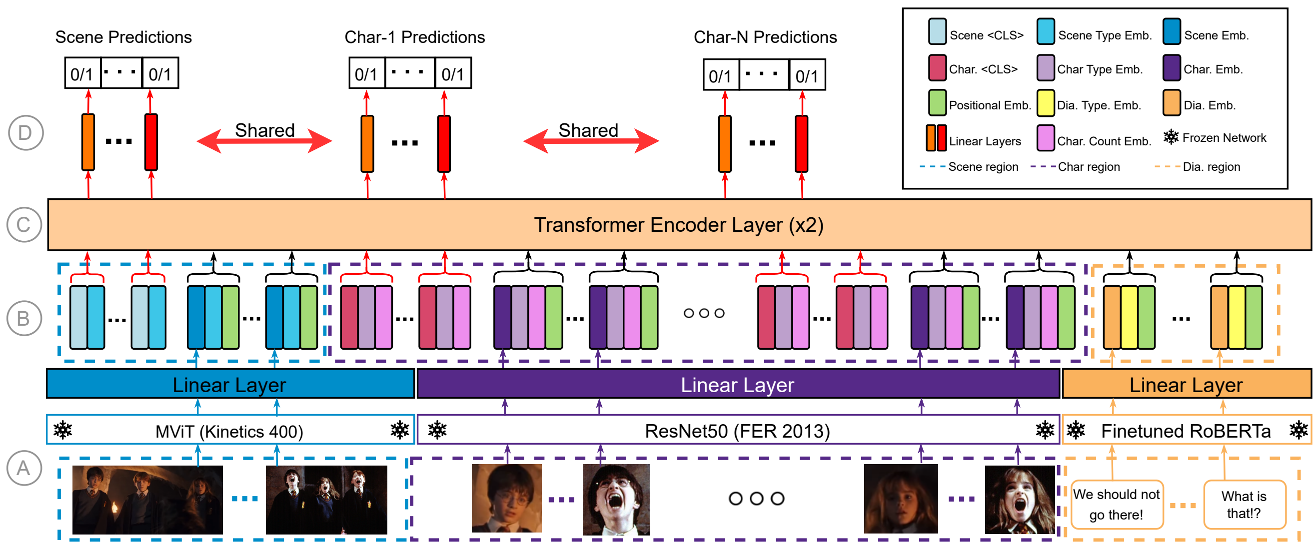

Figure 2. An overview of EmoTx. We present the detailed approach in Sec. 3 but provide a short summary here. **A**: Video features (in blue region), character face features (in purple region), and utterance features (in orange region) are obtained using frozen backbones and projected with linear layers into a joint embedding space. **B**: Here appropriate embeddings are added to the tokens to distinguish between modalities, character count, and to provide a sense of time. We also create per-emotion classifier tokens associated with the scene or a specific character. **C**: Two Transformer encoder layers perform self-attention across the sequence of input tokens. **D**: Finally, we tap the classifier tokens to produce output probability scores for each emotion through a linear classifier shared across the scene and characters.

relevant time bins. For an utterance u_j, binning is done based on its middle timestamp t_j. We denote the time embeddings as $\mathbf{E}^T \in \mathbb{R}^{D\times\lceil T^*/\tau\rceil}$, where T^* is the maximum scene duration and τ is the bin step. For convenience, $\mathbf{E}^T_t$ selects the embedding using a discretized index $\lceil t/\tau \rceil$.

Classifier tokens. Similar to the classic CLS tokens in Transformer models [17, 85] we use learnable classifier tokens to predict the emotions. Furthermore, inspired by Query2Label [40], we use K classifier tokens rather than tapping a single token to generate all outputs (see Fig. 2D). This allows capturing label co-occurrence within the Transformer layers improving performance. It also enables analysis of per-emotion attention scores providing insights into the model's workings. In particular, we use K classifier tokens for scene-level predictions (denoted $\mathbf{z}^{\mathcal{S}}_k$) and $N \times K$ tokens for character-level predictions (denoted $\mathbf{z}^i_k$ for character $\mathcal{P}^i$, one for each character-emotion pair).

Token representations. Combining the features with relevant embeddings provides rich information to EmoTx. The token representations for each input group are as follows:

$$\text{scene cls. tokens: } \tilde{\mathbf{z}}^{\mathcal{S}}_k = \mathbf{z}^{\mathcal{S}}_k + \mathbf{E}^{\mathcal{M}}_1, \quad (1)$$

$$\text{char. cls. tokens: } \tilde{\mathbf{z}}^{i}_k = \mathbf{z}^{i}_k + \mathbf{E}^{\mathcal{M}}_2 + \mathbf{E}^{C}_i, \quad (2)$$

$$\text{video: } \tilde{\mathbf{f}}_t = \mathbf{W}_{\mathcal{V}}\mathbf{f}_t + \mathbf{E}^{\mathcal{M}}_1 + \mathbf{E}^{T}_t, \quad (3)$$

$$\text{character box: } \tilde{\mathbf{c}}^i_t = \mathbf{W}_{\mathcal{C}}\mathbf{c}^i_t + \mathbf{E}^{\mathcal{M}}_2 + \mathbf{E}^{C}_i + \mathbf{E}^{T}_t, \quad (4)$$

$$\text{and utterance: } \tilde{\mathbf{u}}_j = \mathbf{W}_{\mathcal{U}}\mathbf{u}_j + \mathbf{E}^{\mathcal{M}}_3 + \mathbf{E}^{T}_{t_j}. \quad (5)$$

Fig. 2B illustrates this addition of embedding vectors. We also perform LayerNorm [2] before feeding the tokens to the Transformer encoder layers, not shown for brevity.

Transformer Self-attention. We concatenate and pass all tokens through H=2 layers of the Transformer encoder that computes self-attention across all modalities [71]. For emotion prediction, we only tap the outputs corresponding to the classification tokens as

$$[\hat{\mathbf{z}}^{\mathcal{S}}_k, \hat{\mathbf{z}}^i_k] = \mathsf{TransformerEncoder}\left(\tilde{\mathbf{z}}^{\mathcal{S}}_k, \tilde{\mathbf{f}}_t, \tilde{\mathbf{z}}^i_k, \tilde{\mathbf{c}}^i_t, \tilde{\mathbf{u}}_j\right). \quad (6)$$

We jointly encode all tokens spanning $\{k\}^K_1, \{i\}^N_1, \{t\}^T_1$, and $\{j\}^M_1$.

Emotion labeling. The contextualized representations for the scene $\hat{\mathbf{z}}^{\mathcal{S}}_k$ and characters $\hat{\mathbf{z}}^i_k$ are sent to a shared linear layer $\mathbf{W}^E \in \mathbb{R}^{K\times D}$ for classification. Finally, the probability estimates through a sigmoid activation $\sigma(\cdot)$ are:

$$\hat{y}^{\mathcal{S}}_k = \sigma(\mathbf{W}^E_k \hat{\mathbf{z}}^{\mathcal{S}}_k) \text{ and } \hat{y}^i_k = \sigma(\mathbf{W}^E_k \hat{\mathbf{z}}^i_k), \ \forall k, i. \quad (7)$$

3.3. Training and Inference

Training. EmoTx is trained in an end-to-end fashion with the *BinaryCrossEntropy* (BCE) loss. To account for the class imbalance we provide weights ω_k for the positive labels based on inverse of proportions. The scene and character prediction losses are combined as

$$\mathcal{L} = \sum_{k=1}^{K} \mathsf{BCE}(\omega_k, y^{\mathcal{V}}_k, \hat{y}^{\mathcal{S}}_k) + \sum_{i=1}^{N}\sum_{k=1}^{K} \mathsf{BCE}(\omega_k, y^{\mathcal{P}^i}_k, \hat{y}^i_k). \quad (8)$$

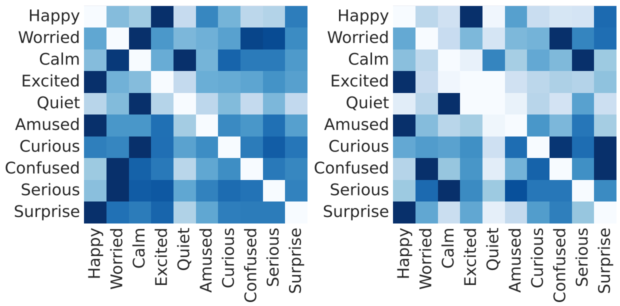

Figure 3. Row normalized label co-occurrence matrices for the top-10 emotions in a *movie scene* (left) or for a *character* (right).

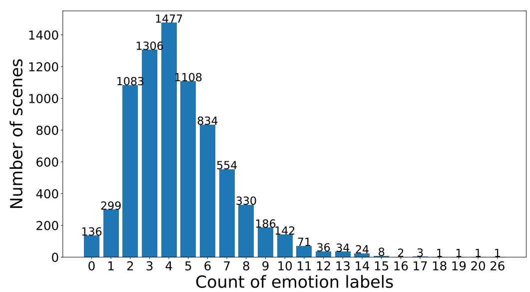

Figure 4. Bar chart showing the number of movie scenes associated with a specific count of annotated emotions.

Inference. At test time, we follow the procedure outlined in Sec. 3.2 and generate emotion label estimates for the entire scene and each character as indicated in Eq. 7.

Variations. As we will see empirically, our model is very versatile and well suited for adding/removing modalities or additional representations by adjusting the width of the Transformer (number of tokens). It can be easily modified to act as a unimodal architecture that applies only to video or dialog utterances by disregarding other modalities.

4. Experiments and Discussion

We present our experimental setup in Sec. 4.1 before diving into the implementation details in Sec. 4.2. A series of ablation studies motivate the design choices of our model (Sec. 4.3) while we compare against the adapted versions of various SoTA models for emotion recognition in Sec. 4.4. Finally, we present some qualitative analysis and discuss how our model switches from facial expressions to video or dialog context depending on the label in Sec. 4.5.

4.1. Dataset and Setup

We use the MovieGraphs dataset [72] that features 51 movies and 7637 movie scenes with detailed graph annotations. We focus on the list of characters and their emotions and mental states, which naturally affords a multi-label setup. Other annotations such as the situation label, or character interactions and relationships [32] are ignored as they cannot be assumed to be available for a new movie.

Label sets. Like other annotations in the MovieGraphs dataset, emotions are also obtained as free-text leading to a huge variability and a long-tail of labels (over 500). We focus our experiments on three types of label sets: (i) *Top-10* considers the most frequently occurring 10 emotions; (ii) *Top-25* considers frequently occurring 25 labels; and (iii) *Emotic*, a mapping from 181 MovieGraphs emotions to 26 Emotic labels provided by [45].

Statistics. We first present row max-normalized co-occurrence matrices for the scene and characters (Fig. 3). It is interesting to note how a movie scene has high co-occurrence scores for emotions such as *worried* and *calm* (perhaps owing to multiple characters), while *worried* is most associated with *confused* for a single character. Another high scoring example for a single character is *curious* and *surprise*, while a movie scene has *curious* with *calm* and *surprise* with *happy*. In Fig. 4, we show the number of movie scenes that contain a specified number of emotions. Most scenes have 4 emotions. The supplementary material section B features further analysis.

Evaluation metric. We use the original splits from MovieGraphs. As we have K binary classification problems, we adopt mean Average Precision (mAP) to measure model performance (similar to Atomic Visual Actions [25]). Note that AP also depends on the label frequency.

4.2. Implementation Details

Feature representations play a major role on the performance of any model. We describe different backbones used to extract features for video frames, characters, and dialog.

Video features $\mathbf{f}_t$: The visual context is important for understanding emotions [31, 34, 44]. We extract spatial features using ResNet152 [26] trained on ImageNet [59], ResNet50 [26] trained on Place365 [82], and spatio-temporal features, MViT [20] trained on Kinetics400 [10].

Dialog features $\mathbf{u}_j$: Each utterance is passed through a RoBERTa-Base encoder [85] to obtain an utterance-level embedding. We also extract features from a RoBERTa model fine-tuned for the task of multi-label emotion classification (based on dialog only).

Character features $\mathbf{c}_t^i$: are represented based on face or person detections. We perform face detection with MTCNN [81] and person detection with Cascade RCNN [8] trained on MovieNet [27]. Tracks are obtained using SORT [5], a simple Kalman filter based algorithm, and clusters using C1C [29]. Details of the character processing pipeline are presented in the supplement section C. ResNet50 [1] trained on SFEW [14] and pretrained on FER13 [24] and VGGFace [51], VGGm [1] trained on FER13 and pretrained on VGGFace, and InceptionResnetV1 [64] trained on VGGFace2 [9] are used to extract face representations.

Method	Top-10		Top-25	
	Scene	Char	Scene	Char
Random	16.87±0.23	12.49±0.15	9.73±0.101	5.84±0.05
MLP (2 Lin)	23.94±0.03	20.39±0.01	15.26±0.02	10.57±0.02
Single Tx encoder	25.66±0.02	20.95±0.09	16.14±0.03	11.08±0.18
EmoTx: 1 CLS	*34.11*±0.34	*23.81*±0.24	*23.34*±0.11	12.86±0.11
EmoTx (Ours)	**34.22**±0.18	**24.35**±0.23	**23.86**±0.10	**13.36**±0.11

Table 1. Architecture ablation. Emotions are predicted at both movie scene and individual character (Char) levels. We see that our multimodal model significantly outperforms simpler baselines. Best numbers in bold, close second in italics.

	V_r	V_m	D	C	Top 10 (mAP)		Top 25 (mAP)	
					Scene	Char	Scene	Char
1	✓	-	-	-	22.81±0.02	15.90±0.19	14.85±0.02	7.98±0.05
2	-	✓	-	-	25.73±0.02	17.88±0.12	16.11±0.05	8.96±0.12
3	-	-	✓	-	27.28±0.01	20.25±0.14	20.20±0.08	11.09±0.12
4	-	-	-	✓	31.38±0.40	21.22±0.50	20.32±0.05	11.23±0.14
5	✓	-	✓	-	27.19±0.07	19.45±0.10	19.72±0.03	10.67±0.08
6	-	✓	✓	-	28.93±0.02	21.41±0.15	21.29±0.05	12.03±0.23
7	-	-	✓	✓	33.59±0.10	23.54±0.16	23.40±0.09	13.01±0.08
8	✓	-	✓	✓	33.60±0.02	22.89±0.02	22.76±0.02	12.21±0.02
9	-	✓	✓	✓	**34.22** ±0.18	**24.35**±0.23	**23.86**±0.10	**13.36**±0.11

Table 2. Modality ablation. V_r: ResNet50 (Places365), V_m: MViT (Kinetics400), D: Dialog, and C: Character.

Frame sampling strategy. We sample up to T=300 tokens at 3 fps (100 s) for the video modality. This covers ∼99% of all movie scenes. Our time embedding bins are also at 3 per second, *i.e.* τ=0.333 s. During inference, a fixed set of frames are chosen, while during training, frames are randomly sampled from 3 fps intervals which acts as data augmentation. Character tokens are treated in a similar fashion, however are subject to the character appearing in the video.

Architecture details. We experiment with the number of encoder layers, $H \in \{1, 2, 4, 8\}$, but find H=2 to work best (perhaps due to the limited size of the dataset). Both the layers have same configuration - 8 attention heads with hidden dimension of 512. The maximum number of characters is N=4 as it covers up to 91% of the scenes. Tokens are padded to create batches and to accommodate shorter video clips. Appropriate masking prevents self-attention on padded tokens. Put together, EmoTx encoder looks at K scene classification tokens, T video tokens, $N \cdot (K + T)$ character tokens, and T utterance tokens. For K=25, N=4 (Top-25 label set), this is up to 1925 padded tokens.

Training details. Our model is implemented in PyTorch [52] and trained on a single NVIDIA GeForce RTX-2080 Ti GPU for a maximum of 50 epochs with a batch size of 8. The hyperparameters are tuned to achieve best performance on validation set. We adopt the Adam optimizer [30] with an initial learning rate of 5×10^{-5}, reduced by a factor of 10 using the learning rate scheduler `ReduceLROnPlateau`. The best checkpoint maximizes the geometric mean of scene and character mAP.

4.3. Ablation Studies

We perform ablations across three main dimensions: architectures, modalities, and feature backbones. When not mentioned, we adopt the defaults: (i) MViT trained on Kinetics400 dataset to represent video; (ii) ResNet50 trained on SFEW, FER, and VGGFace for character representations; (iii) fine-tuned RoBERTa for dialog utterance representations; and (iv) EmoTx with appropriate masking to pick modalities or change the number of classifier tokens.

Architecture ablations. We compare our architecture against simpler variants in Table 1. The first row sets the expectation by providing scores for a *random* baseline that samples label probabilities from a uniform random distribution between $[0, 1]$ with 100 trials. Next, we evaluate *MLP (2 Lin)*, a simple MLP with two linear layers with inputs as max pooled scene or character features. An alternative to max pooling is self-attention. The *Single Tx encoder* performs self-attention over features (as tokens) and a classifier token to which a multi-label classifier is attached. Both these approaches are significantly better than random, especially for individual character level predictions which are naturally more challenging than scene-level predictions.

Finally, we compare multimodal EmoTx that uses 1 classifier token to predict all labels (EmoTx: 1 CLS) against K classifier tokens (last row). Both models achieve significant improvements, *e.g.* in absolute points, +8.5% for Top-10 scene labels and +2.3% for the much harder Top-25 character level labels. We believe the improvements reflect EmoTx's ability to encode multiple modalities in a meaningful way. Additionally, the variant with K classifier tokens (last row) shows small but consistent +0.5% improvements over 1 classifier token on Top-25 emotions.

Fig. 5 shows the scene-level AP scores for the Top-25 labels. Our model outperforms the MLP and Single Tx encoder on 24 of 25 labels and outperforms the single classifier token variant on 15 of 25 labels. EmoTx is good at recognizing expressive emotions such as *excited, serious, happy* and even mental states such as *friendly, polite, worried*. However, other mental states such as *determined* or *helpful* are challenging.

Modality ablations. We evaluate the impact of each modality (video, characters, and utterances) on scene- and character-level emotion prediction in Table 2. We observe that the character modality (row 4, R4) outperforms any of the video or dialog modalities (R1-R3). Similarly, dialog features (R3) are better than video features (R1, R2), common in movie understanding tasks [68, 72]. The choice of

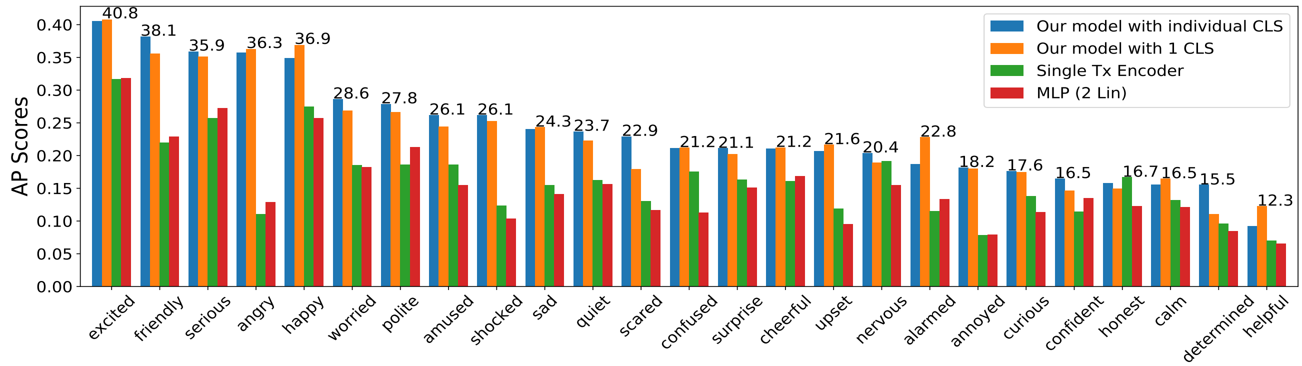

Figure 5. Comparing scene-level per class AP of EmoTx against baselines (Table 1) shows consistent improvements. We also see that our model with K classifier tokens outperforms the 1 CLS token on most classes. AP of the best model is indicated above the bar. Interestingly, the order in which emotions are presented is not the same as the frequency of occurrence (see supplement section B).

	Video		Character		Dialog	Metrics			
	MViT	R50	R50	VGG-M	RB	Top-10		Top-25	
	K400	P365	FER	FER	FT	Scene	Char	Scene	Char
1	-	✓	-	✓	No	29.30	19.73	19.05	10.31
2	✓	-	-	✓	No	29.34	20.50	19.07	10.34
3	-	✓	✓	-	No	29.69	20.25	20.16	11.06
4	✓	-	✓	-	No	31.39	21.12	20.88	11.46
5	✓	-	-	✓	✓	31.50	21.60	21.49	11.64
6	-	✓	-	✓	✓	32.42	22.32	21.45	11.62
7	-	✓	✓	-	✓	33.46	22.98	22.69	12.48
8	✓	-	✓	-	✓	**34.22**	**24.35**	**23.86**	**13.36**

Table 3. Feature ablations with backbones. (MViT, K400): MViT on Kinetics400, (R50, P365): ResNet50 on Places365, (R50, FER): ResNet50 on Facial Expression Recognition (FER), (VGG-M, FER): VGG-M on FER, and (RB, FT): RoBERTa finetuned. Best numbers in bold. More results in supplement E.

Method	Top 10		Top 25		Emotic	
	Val	Test	Val	Test	Val	Test
Random	16.87	13.84	9.73	7.57	11.47	11.36
CAER [34]	18.35	15.38	11.84	9.49	13.91	12.68
ENet [73]	19.14	16.14	11.22	9.08	13.55	12.64
AANet [69]	21.55	17.55	12.55	10.20	14.71	13.37
M2Fnet [12]	24.55	19.10	16.02	13.05	18.27	16.76
EmoTx (Ours)	**34.22**	**29.35**	**23.86**	**19.47**	**23.67**	**21.40**

Table 4. Comparison against SoTA for scene-level predictions. *AANet*: AttendAffectNet. *ENet*: EmotionNet. Mean over 3 runs.

Method	Top 10		Top 25		Emotic	
	Val	Test	Val	Test	Val	Test
Random	12.49	11.37	5.84	5.36	6.40	6.32
AANet [69]	17.43	16.04	8.64	7.20	8.53	7.75
M2Fnet [12]	20.82	19.01	10.67	9.71	11.30	9.92
EmoTx (Ours)	**24.35**	**22.32**	**13.36**	**11.71**	**12.29**	**11.76**

Table 5. Comparison against SoTA for character-level predictions. *AANet* denotes AttendAffectNet. Mean over 3 runs.

visual features is important. Scene features V_r are consistently worse than action features V_m as reflected in comparisons R1, R2 or R5, R6 or R8, R9. Finally, we observe that using all modalities (R9) outperforms other combinations, indicating that emotion recognition is a multimodal task.

Backbone ablations. We compare several backbones for the task of emotion recognition. The effectiveness of the fine-tuned RoBERTa model is evident by comparing pairs of rows R2, R5 and R3, R7 and R4, R8 of Table 3, where we see a consistent improvement of 1-3%. Character representations with ResNet50-FER show improvement over VGGm-FER as seen from R5, R8 or R6, R7. Finally, comparing R8 shows the benefits provided by action features as compared to places. Detailed results are presented in the supplement, section E.

4.4. SoTA Comparison

We compare our model against published works EmotionNet [73], CAER [34], AttendAffectNet [69], and M2Fnet [12] by adapting them for our tasks (adaptation details are provided in the supplement, section F). Table 4 shows scene-level performance while the character-level performance is presented in Table 5. First, we note that the test set seems to be harder than val as also indicated by the random baseline, leading to a performance drop from val to test across all approaches. EmoTx outperforms all previous baselines by a healthy margin. For scene level, we see +4.6% improvement on Emotic labels, +7.8% on Top-25, and +9.7% on Top-10. Character-level predictions are more challenging, but we see consistent improvements of +1.5-3% across all label sets. Matching expectation, we see that simpler models such as EmotionNet or CAER perform worse than Transformer-based approaches of M2Fnet and AttendAffectNet. Note that EmotionNet and CAER are challenging to adapt for character-level predictions and are not presented, but we expect M2Fnet or AttendAffectNet to outperform them.

Figure 6. A scene from the movie *Forrest Gump* showing the multimodal self-attention scores for the two predictions: *Mrs. Gump* is *worried* and *Forrest* is *happy*. We observe that the *worried* classifier token attends to *Mrs. Gump*'s character tokens when she appears at the start of the scene, while *Forrest*'s *happy* classifier token attends to *Forrest* towards the end of the scene. The video frames have relatively similar attention scores while dialog helps with emotional utterances such as *told you not to bother* or *it sounded good*.

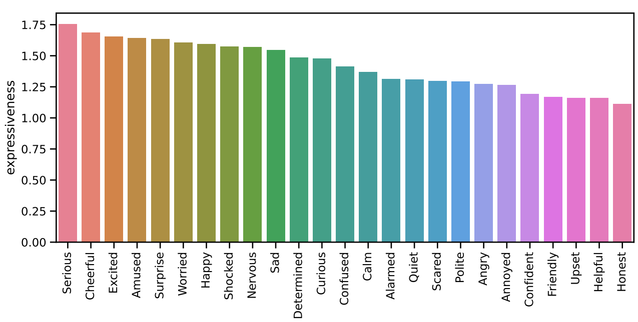

Figure 7. Sorted expressiveness scores for Top-25 emotions. Expressive emotions have higher scores indicating that the model attends to character representations, while mental states have lower scores suggesting more attention to video and dialog context.

4.5. Analyzing Self-attention Scores

EmoTx provides an intuitive way to understand which modalities are used to make predictions. We refer to the self-attention scores matrix as α, and analyze specific rows and columns. Separating the K classifier tokens allows us to find attention-score based evidence for each predicted emotion by looking at a row $\alpha_{\mathbf{z}_k^{\mathcal{S}}}$ in the matrix.

Fig. 6 shows an example movie scene where EmoTx predicts that *Forrest* is *happy* and *Mrs. Gump* is *worried*. We see that the model pays attention to the appropriate moments and modalities to make the right predictions.

Expressive emotions *vs*. Mental states. We hypothesize that the self-attention module may focus on character tokens for expressive emotions, while looking at the overall video frames and dialog for the more abstract mental states. We propose an *expressiveness* score as

$$e_k = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \alpha_{\mathbf{z}_k^{\mathcal{S}}, \mathbf{c}_t^i}}{\sum_{t=1}^{T} \alpha_{\mathbf{z}_k^{\mathcal{S}}, \mathbf{f}_t} + \sum_{j=1}^{M} \alpha_{\mathbf{z}_k^{\mathcal{S}}, \mathbf{u}_j}}, \tag{9}$$

where $\alpha_{\mathbf{z}_k^{\mathcal{S}}, \mathbf{c}_t^i}$ is the self-attention score between the scene classifier token for emotion k ($\mathbf{z}_k^{\mathcal{S}}$) and character $\mathcal{P}^i$'s appearance in the video frame as b_t^i; $\alpha_{\mathbf{z}_k^{\mathcal{S}}, \mathbf{f}_t}$ is for the video f_t and $\alpha_{\mathbf{z}_k^{\mathcal{S}}, \mathbf{u}_j}$ is for dialog utterance u_j. Higher scores indicate expressive emotions as the model focuses on the character features, while lower scores identify mental states that analyze the video and dialog context. Fig. 7 shows the averaged expressiveness score for the Top-25 emotions when the emotion is present in the scene (*i.e.* y_k=1). We observe that mental states such as *honest, helpful, friendly, confident* appear towards the latter half of this plot while most expressive emotions such as *cheerful, excited, serious, surprise* appear in the first half. Note that the expressiveness scores in our work are for faces and applicable to our particular dataset. We also conduct a short human evaluation to understand expressiveness by annotating whether the emotion is conveyed through video, dialog, or character appearance; presented in the supplement section G.

5. Conclusion

We presented a novel task for multi-label emotion and mental state recognition at the level of a movie scene and for each character. A Transformer encoder based model, EmoTx, was proposed that jointly attended to all modalities (features) and obtained significant improvements over previous works adapted for this task. Our learned model was shown to have interpretable attention scores across modalities – they focused on the video or dialog context for mental states while looking at characters for expressive emotions. In the future, EmoTx may benefit from audio features or by considering the larger context of the movies instead of treating every scene independently.

Acknowledgements. We thank Google India Faculty Research Award 2022 for travel support.

References

[1] S. Albanie and A. Vedaldi. Learning Grimaces by Watching TV. In *British Machine Vision Conference (BMVC)*, 2016. 5

[2] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *arXiv: 1607.06450*, 2016. 4

[3] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed Movies: Story Based Retrieval with Contextual Embeddings. In *Asian Conference on Computer Vision (ACCV)*, 2020. 2

[4] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, pages 43–55, 2015. 1, 2

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple Online and Realtime Tracking. In *International Conference on Image Processing (ICIP)*, 2016. 5

[6] Andrew Brown, Ernesto Coto, and Andrew Zisserman. Automated Video Labelling: Identifying Faces by Corroborative Evidence. In *Multimedia Information Processing and Retrieval (MIPR)*, 2021. 2

[7] Andrew Brown, Vicky Kalogeiton, and Andrew Zisserman. Face, Body, Voice: Video Person-Clustering with Multiple Modalities. In *International Conference on Computer Vision Workshops (ICCVW)*, 2021. 2

[8] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into High Quality Object Detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[9] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *International Conference on Automatic Face and Gesture Recognition (FG)*, 2018. 5

[10] João Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[11] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Self-Supervised Learning for Scene Boundary Detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[12] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 2, 7

[13] Gerald L. Clore, Andrew Ortony, and Mark A. Foss. The Psychological Foundations of the Affective Lexicon. *Journal of Personality and Social Psychology*, 53(4):751–766, 1987. 1, 2

[14] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *International Conference on Computer Vision Workshops (ICCVW)*, 2011. 5

[15] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE Multimedia*, 19:34–41, 2012. 2

[16] Dhall, Abhinav and Goecke, Roland and Joshi, Jyoti and Wagner, Michael and Gedeon, Tom. Emotion recognition in the wild challenge 2013. In *International Conference on Multimodal Interaction (ICMI)*, 2013. 2

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021. 4

[18] Paul Ekman and W V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, pages 124–9, 1971. 1, 2

[19] Mark Everingham, Josef Sivic, and Andrew Zisserman. "Hello! My name is ... Buffy" – Automatic Naming of Characters in TV Video. In *British Machine Vision Conference (BMVC)*, 2006. 2, 3

[20] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale Vision Transformers. In *International Conference on Computer Vision (ICCV)*, 2021. 5

[21] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding Human Gaze Communication by Spatio-Temporal Graph Reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[22] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Findings of Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 2

[23] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 2

[24] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing (ICONIPS)*, 2013. 2, 5

[25] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 5

[27] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. MovieNet: A Holistic Dataset for Movie Understanding. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 5

[28] Wenxiang Jiao, Michael Lyu, and Irwin King. Real-Time Emotion Recognition via Attention Gated Hierarchical Memory Network. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020. 2

[29] Kalogeiton, Vicky, and Zisserman, Andrew. Constrained video face clustering using 1nn relations. In *British Machine Vision Conference (BMVC)*, 2020. 2, 5

[30] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations (ICLR)*, 2015. 6

[31] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 5

[32] Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. Learning Interactions and Relationships between Movie Characters. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5

[33] Joseph E. LeDoux. Evolution of Human Emotions. *Progress in Brain Research*, 195:431–442, 2013. 1

[34] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware Emotion Recognition Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 5, 7

[35] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, Compositional Video Question Answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 2

[36] Juncheng Li, Junlin Xie, Linchao Zhu, Long Qian, Siliang Tang, Wenqiao Zhang, Haochen Shi, Shengyu Zhang, Longhui Wei, Qi Tian, and Yueting Zhuang. Dilated Context Integrated Network with Cross-Modal Consensus for Temporal Emotion Localization in Videos. In *ACM Multimedia (MM)*, 2022. 3

[37] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation. In *Advances in Neural Information Processing Systems (NeurIPS): Track on Datasets and Benchmarks*, 2021. 2

[38] Mengyi Liu, Shaoxin Li, S. Shan, Ruiping Wang, and Xilin Chen. Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis. In *Asian Conference on Computer Vision (ACCV)*, 2014. 2

[39] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial Expression Recognition via a Boosted Deep Belief Network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1805–1812, 2014. 2

[40] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2Label: A Simple Transformer Way to Multi-Label Classification. *arXiv:2107.10834*, 2021. 3, 4

[41] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason M. Saragih, Zara Ambadar, and I. Matthews. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 94–101, 2010. 2

[42] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019. 1, 2

[43] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. LAEO-Net: revisiting people Looking At Each Other in videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[44] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. EmotiCon: Context-Aware Multimodal Emotion Recognition using Frege's Principle. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 5

[45] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. Affect2MM: Affective Analysis of Multimedia Content Using Emotion Causality. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 5

[46] Arsha Nagrani and Andrew Zisserman. From Benedict Cumberbatch to Sherlock Holmes: Character Identification in TV series without a Script. In *British Machine Vision Conference (BMVC)*, 2017. 2

[47] Desmond Ong, Zhengxuan Wu, Tan Zhi-Xuan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. Modeling Emotion in Complex Stories: The Stanford Emotional Narratives Dataset. *IEEE Transactions on Affective Computing*, 2019. 2

[48] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K. Roy-Chowdhury. Contemplating Visual Emotions: Understanding and Overcoming Dataset Bias. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *European Conference on Computer Vision (ECCV)*, 2018. 2

[49] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *International Conference on Multimedia and Expo (ICME)*, 2005. 2

[50] Jae Sung Park, Trevor Darrell, and Anna Rohrbach. Identity-Aware Multi-Sentence Video Description. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[51] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep Face Recognition. In *British Machine Vision Conference (BMVC)*, 2015. 5

[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 6

[53] Robert Plutchik. A General Pscychoevolutionary Theory of Emotion. *Theories of Emotion*, pages 3–33, 1980. 1
[54] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Association of Computational Linguistics (ACL)*, 2019. 1, 2
[55] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A Local-to-Global Approach to Multi-modal Movie Scene Segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3
[56] Zeeshan Rasheed and Mubarak Shah. Scene Detection in Hollywood Movies and TV Shows. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 2
[57] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie Description. *IJCV*, 123:94–120, 2017. 2
[58] Daniel Rotman, Dror Porat, and Gal Ashour. Optimal Sequential Grouping for Robust Video Scene Detection using Multiple Modalities. *International Journal of Semantic Computing*, 11(2):192–208, 2017. 2
[59] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115:211–252, 2015. 5
[60] Amy M. Schmitter. 17th and 18th Century Theories of Emotions. In *The Stanford Encyclopedia of Philosophy*, 2021. 1
[61] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2021. 2
[62] Sarath Sivaprasad, Tanmayee Joshi, Rishabh Agrawal, and Niranjan Pedanekar. Multimodal Continuous Prediction of Emotions in Movies using Long Short-Term Memory Networks. In *International Conference on Multimedia Retrieval (ICMR)*, 2018. 2
[63] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A Joint Model for Video and Language Representation Learning. In *International Conference on Computer Vision (ICCV)*, 2019. 3
[64] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5
[65] Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. "Knock! Knock! Who is it?" Probabilistic Person Identification in TV series. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2
[66] Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. StoryGraphs: Visualizing Character Interactions as a Timeline. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3
[67] Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. Book2Movie: Aligning Video scenes with Book chapters. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2
[68] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 6
[69] Ha Thi Phuong Thao, BT Balamurali, Dorien Herremans, and Gemma Roig. AttendAffectNet: Self-Attention based Networks for Predicting Affective Responses from Movies. In *International Conference on Pattern Recognition (ICPR)*, 2021. 2, 7
[70] Y.-I. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(2):97–115, 2001. 2
[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3, 4
[72] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. MovieGraphs: Towards Understanding Human-Centric Situations from Videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 5, 6
[73] Zijun Wei, Jianming Zhang, Zhe Lin, Joon-Young Lee, Niranjan Balasubramanian, Minh Hoai, and Dimitris Samaras. Learning Visual Emotion Representations From Web Data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 7
[74] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth S. Narayanan. Context-sensitive Multimodal Emotion Recognition from Speech and Facial Expression using Bidirectional LSTM Modeling. In *Interspeech*, 2010. 2
[75] Chao-Yuan Wu and Philipp Krähenbühl. Towards Long-Form Video Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
[76] Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. A Graph-based Framework to Bridge Movies and Synopses. In *International Conference on Computer Vision (ICCV)*, 2019. 2
[77] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A Joint Sequence Fusion Model for Video Question Answering and Retrieval. In *European Conference on Computer Vision (ECCV)*, 2018. 2
[78] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end Concept Word Detection for Video Captioning, Retrieval, and Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2
[79] Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Multi-modal Multi-label Emotion Detection with Modality and Label Dependence. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 3

[80] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019. 2

[81] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, pages 1499–1503, 2016. 5

[82] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(6):1452–1464, 2017. 5

[83] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In *International Joint Conference on Natural Language Processing (IJCNLP)*, 2021. 1, 3

[84] Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[85] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Chinese National Conference on Computational Linguistics*, 2021. 4, 5