Dilated Context Integrated Network with Cross-Modal Consensus for Temporal Emotion Localization in Videos

Juncheng Li*
Zhejiang University
junchengli@zju.edu.cn

Junlin Xie*
Zhejiang University
junlinxie@zju.edu.cn

Linchao Zhu
University of Technology Sydney
linchao.zhu@uts.edu.au

Long Qian
Zhejiang University
qianlong0926@zju.edu.cn

Siliang Tang
Zhejiang University
siliang@zju.edu.cn

Wenqiao Zhang†
National University of Singapore
wenqiao@nus.edu.sg

Haochen Shi
Université de Montréal
haochen.shi@umontreal.ca

Shengyu Zhang
Zhejiang University
sy_zhang@zju.edu.cn

Longhui Wei
Huawei Cloud
weilh2568@gmail.com

Qi Tian
Huawei Cloud
tian.qi1@huawei.com

Yueting Zhuang
Zhejiang University
yzhuang@zju.edu.cn

ABSTRACT

Understanding human emotions is a crucial ability for intelligent robots to provide better human-robot interactions. The existing works are limited to trimmed video-level emotion classification, failing to locate the temporal window corresponding to the emotion. In this paper, we introduce a new task, named Temporal Emotion Localization in videos (TEL), which aims to detect human emotions and localize their corresponding temporal boundaries in untrimmed videos with aligned subtitles. TEL presents three unique challenges compared to temporal action localization: 1) The emotions have extremely varied temporal dynamics; 2) The emotion cues are embedded in both appearances and complex plots; 3) The fine-grained temporal annotations are complicated and labor-intensive. To address the first two challenges, we propose a novel dilated context integrated network with a coarse-fine two-stream architecture. The coarse stream captures varied temporal dynamics by modeling multi-granularity temporal contexts. The fine stream achieves complex plots understanding by reasoning the dependency between the multi-granularity temporal contexts from the coarse stream and adaptively integrates them into fine-grained video segment features. To address the third challenge, we introduce a cross-modal consensus learning paradigm, which leverages the inherent semantic consensus between the aligned video and subtitle to achieve weakly-supervised learning. We contribute a new testing set with 3,000 manually-annotated temporal boundaries so that future research on the TEL problem can be quantitatively evaluated. Extensive experiments show the effectiveness of our approach on temporal emotion localization. The repository of this work is at https://github.com/YYJMJC/Temporal-Emotion-Localization-in-Videos.

Figure 1: We show some key segments and their aligned subtitles. The segments in the red box are the target segments that temporally occur *Anger* emotion.

CCS CONCEPTS

• **Information systems** → *Multimedia information systems*; • **Computing methodologies** → *Computer vision.*

KEYWORDS

Weakly-Supervised Temporal Emotion Localization; Cross-Modal Consensus; Video-and-Language Understanding;

ACM Reference Format:
Juncheng Li, Junlin Xie, Linchao Zhu, Long Qian, Siliang Tang, Wenqiao Zhang, Haochen Shi, Shengyu Zhang, Longhui Wei, Qi Tian, and Yueting Zhuang. 2022. Dilated Context Integrated Network with Cross-Modal Consensus for Temporal Emotion Localization in Videos. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Oct. 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547886

* Equal Contribution.
†Corresponding Author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3547886

1 INTRODUCTION

Humans are social creatures and thrive on empathy. We can easily put ourselves in other's situations and make decisions based on the inferences of other's internal states (*i.e.* emotional states). Recent studies [3, 23] on neuroscience confirm that emotional intelligence and cognitive intelligence share many neural systems for integrating cognitive, social, and affective processes. Therefore, understanding emotions is crucial for achieving high-level intelligence. In practice, this ability helps social chatbots and personal assistants to better understand the mood and motivations of people, so they can better interact with people.

Emotion understanding has long been studied in computer vision. Previous studies [2, 14, 55, 62] mainly focus on recognizing emotion through facial expression in static images. As movies provide diverse social situations that are closer to our daily life, some recent studies [21, 51, 52] have been proposed to classify emotions from movies. However, they are limited to video-level classification, failing to identify the corresponding temporal boundaries of the emotions, which is essential in practical application. To break through the above limitation, we propose a novel task of Temporal Emotion Localization in videos (TEL), which aims to predict emotions and their corresponding start and end timestamps in untrimmed videos with aligned subtitles.

Compared with conventional temporal action localization (TAL) [6], TEL presents some unique challenges. First, the emotions have extremely varied temporal dynamics. Such fine-grained emotions could appear in arbitrary frames and last for varied durations, causing a great challenge for temporal localizing. For example, the peace emotion may exist for a long time but the surprise emotion may happen quickly. Even the same emotion may have extremely varied durations in different situations. Secondly, unlike existing action localization, where the actions have more consistent visual patterns and only rely on a single modality (video) as the context, in TEL, the emotion cues are embedded in both appearances and complex plots with multi-modality context. In TEL, the videos paired with subtitles come from movies, which contain diverse event dynamics and character interactions, and the emotions are more ambiguous and have higher inter-class similarity than action classes. To discriminate very similar emotions, the model needs to achieve in-depth comprehension of complex plots by jointly reasoning over multi-modal and multi-granularity temporal context. As shown in Figure 1, when we look only at the target segments in the red box, we can guess that these boys are playing and feeling *Happiness* and *Engagement*, but it is hard to identify more specific cues for their emotions. When we further see the corresponding subtitles, we may infer that they are chasing and feeling *Excitement*. However, only when we consider the whole context that they are bullying a boy and chasing him, can we say they are probably feeling anger. Thirdly, annotating fine-grained temporal boundaries of emotions in videos is complicated and labor-intensive. Thus, a weakly-supervised algorithm for TEL is more widely available.

In this paper, we propose a novel dilated context integrated network with cross-modal consensus learning to address the aforementioned challenges. For the first two challenges, we introduce a Dilated Context Integrated Network (DCIN) that adaptively models multi-granularity temporal dynamics to achieve in-depth understanding of complex plots. Specifically, DCIN models temporal context in a coarse-fine two-stream architecture. The coarse stream models multiple abstract-level of context in a hierarchical structure to capture varied temporal dynamics. The fine stream reasons the temporal dependency between the multi-granularity temporal context and adaptively integrates them into fine-grained video segment features by joint reasoning over video and subtitle, which achieves in-depth plots understanding. Furthermore, we present a context-sensitive constraint to encourage the DCIN to learn more discriminative context that can help to determine the emotion.

For weakly-supervised learning, we propose a Cross-Modal Consensus Learning (CCL) paradigm by leveraging the inherent semantic consensus between the aligned video and subtitle. The intuition behind this is that when we see the subtitle "thanks for your delicate gift" we can easily infer the visual situation that somebody is happy and vice versa. If we see a video segment of somebody happy to accept a gift, we may infer some subtitles expressing his happiness. Concretely, given the ground-truth emotion label without temporal annotation, the model first identifies the most relevant video segment and then uses its temporally co-occurring subtitle to predict the most possible emotion. We train our model such that the predicted emotion based on subtitle is consistent with the original emotion for retrieving the most relevant video segment. Further, our empirical experiments indicate that sometimes the alignment between video and subtitle is noisy as the subtitle might refer to previous or forthcoming visual events. Therefore, to alleviate the misalignment noise, we present a temporal alignment relaxation strategy, which enables the model to dynamically learn the alignment from CCL paradigm.

To facilitate research of the TEL task, we contribute a testing set by manually annotating the temporal boundaries of 3000 samples on the MovieGraph dataset [60].

2 RELATED WORK

Emotion Understanding. Emotion understanding has long been studied in computer vision. Existing researchers mainly focus on recognizing emotion through facial expressions. Quiroz *et al.* [14] propose a large dataset of one million images of facial expressions of emotion in the wild. Wei *et al.* [62] perform emotion recognition by learning a feature extraction network on StockEmotion, which has more than a million images. Instead of recognizing emotions only based on facial expressions, there has been a growing interest in dynamically modeling emotions over time. Movies serve as an appropriate testbed of emotion understanding, as they contain multimodal context and diverse human emotions in a variety of situations. Several works [21, 51, 52, 56] have been proposed to classify emotions from movie clips. Although they have achieved promising performance, they still remain in the emotion classification on the whole video, lacking transparency to tell which segments of the video the emotion appears in. In contrast, we further explore localizing the start and end points of emotions in untrimmed videos, which is more challenging and crucial for understanding human emotions in real-world situations (*e.g.*, personal assistants, e-commerce [67]).

Action Localization. Action localization [6] aims to predict actions and corresponding start and end timestamps in videos. In general, existing supervised methods can be categorized into top-down and bottom-up frameworks. The top-down methods [5, 8, 13, 22, 63] first extract a set of candidate proposals and refine them to achieve the final temporal boundaries. The bottom-up methods [4, 45, 46, 48, 49] directly predict frame-level or snippet-level scores and then combine the individual scores to generate the final temporal boundaries. Since supervised action localization requires labor-intensive frame-level annotations, weakly-supervised action localization [50, 53, 54, 61] has received increasing attention.

Video-and-Language Understanding. The advent of deep learning [18, 19, 28, 32, 42] promotes the prosperity of computer vision [26, 41, 75] and vision-and-language [35–37, 39, 65, 69–71, 73]. With the flourishing development of large-scale video datasets [1, 27, 30], several video-and-language understanding tasks [26, 38, 68] have received increasing attention, such as temporal sentence grounding [16, 40, 43, 44], video question answering [33, 57, 72], and video captioning [74]. These tasks mainly focus on identifying explicit visual cues (*e.g.,* objects, actions, characters), which are mainly embedded in obvious visual appearances. TEL differs as it requires more sophisticated reasoning skills, such as understanding complex plots, reasoning character relationships, and inferring human's internal states. These abilities can facilitate more sensible human-robot interactions based on a better comprehension of human emotions.

3 METHOD

Problem Formulation. Given an untrimmed video V paired with subtitle S, we aim to detect emotions in the video and locate their corresponding segments. As an untrimmed video might involve multiple emotions, we formulate the problem as a multi-label detection problem. For weakly-supervised setting, only the video-level emotion labels are available, without any temporal boundary annotations. The video is represented as $V = \{v_i\}_{i=1}^{T}$ segment-by-segment, and the subtitle is represented as $S = \{s_i\}_{i=1}^{T}$ sentence-by-sentence, where s_i represents the subtitle sentence that is temporally co-occurring with v_i. We obtain $v_i \in R^{1\times d}$ by max-pooling over the I3D [7] features of frames within the segment. $s_i \in R^{1\times d}$ is the sentence embedding.

3.1 Dilated Context Integrated Network

As aforementioned, the key factor for fine-grained emotion localization is the multi-modality and multi-granularity context modeling. Existing action localization methods mainly use RNN [24], 3D CNN [7, 58], or Transformer [59] to recognize specific visual patterns of actions. However, they are unsuitable for the complex multi-modality and multi-granularity context modeling. For RNN-based methods, they do not capture non-sequential temporal dependencies effectively. For 3D CNN-based methods, they suffer from limited temporal receptive field. For Transformer-based methods, such fully-connected structures may cause the fine-grained local context to be overwhelmed by unimportant information.

Differently, we present the Dilated Context Integrated Network (DCIN) that adaptively integrates multiple abstract-level of context into fine segment representations in a hierarchy. As shown in Figure 2, DCIN processes the information in a two-stream architecture. The coarse stream models multi-granularity temporal context in a hierarchy. The fine stream reasons the temporal dependency among the temporal context and gradually fuses the multi-granularity context from the coarse stream with the fine segment representations. To avoid unnecessary and redundant context, we propose a gated temporal context integration module to dynamically integrate informative context by joint reasoning over videos and subtitles. Further, we introduce the context-sensitive constraint to encourage the model to learn more discriminative context that can help to determine the emotion. Concretely, each DCIN layer consists of: 1) Temporal Context Convolution, 2) Temporal Context Dependency Reasoning, and 3) Gated Temporal Context Integration.

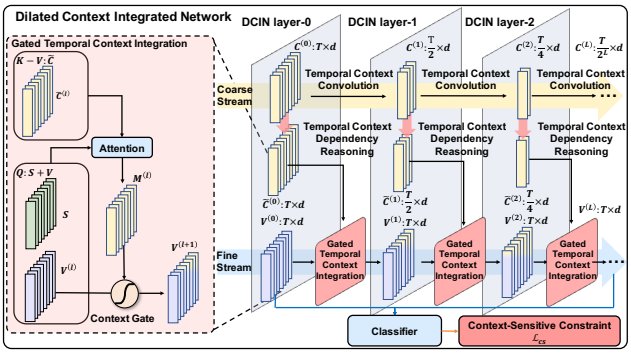

Figure 2: Overview of Dilated Context Integrated Network.

Temporal Context Convolution. For the fine-grained emotion localization, the model must discriminate very similar emotions through video context, which may have various durations and scales. Thus, we use temporal context convolution to generate multi-granularity temporal context. Specifically, we use 1D temporal convolution operation with $stride = 2$ to halve the temporal dimension of the context at each layer. We define the context produced by the Coarse stream at layer l as $C^{(l)}$, where $C^{(0)} = V$. Given $C^{(l-1)} \in R^{T_{l-1}\times d}$ from the previous layer, we compute $C^{(l)} \in R^{T_l\times d}$ ($T_l = T_{l-1}/2$) as:

$$C^{(l)} = f(W_1 * C^{(l-1)} + b_1) \tag{1}$$

where $f(\cdot)$ is the activation function, $*$ is the convolution operator, and W_1 is the 1D convolution filters. As a consequence, we get $C^{(l)}$ that contains increasing levels of semantic meaning and higher temporal resolution context.

Temporal Context Dependency Reasoning. The video clips are collected from movies, which contain complex event dynamics and diverse character interactions across multiple segments. Therefore, we develop the temporal context dependency reasoning to capture the non-local temporal structure of context. Concretely, we adopt graph convolution on the context features $C^{(l)} = \{c_i^l\}_{i=1}^{T_l}$ as:

$$\tilde{c}_i^l = c_i^l + \sum_j^{T_l} \alpha_{ij}^{tcdr} \cdot (W_2 c_j^l), \quad \alpha_{ij}^{tcdr} = \frac{exp({c_i^l}^T \cdot c_j^l)}{\sum_j^{T_l} exp({c_i^l}^T \cdot c_j^l)} \tag{2}$$

where $W_2 \in R^{d\times d}$ is the learnable projection matrix, α_{ij}^{tcdr} is the semantic coefficient between context node c_i^l and c_j^l.

Gated Temporal Context Integration. After obtaining the non-local enhanced context features $\tilde{C}^{(l)}$, we propose a gated temporal context integration module to adaptively integrate context $\tilde{C}^{(l)}$ into segment features $V^{(l-1)}$ at layer $l-1$ ($V^{(0)} = V$) to get $V^{(l)}$. Considering the complementary nature of video and subtitle, we design a cross-modal context filtering mechanism that utilizes the aligned subtitle feature s_i of v_i to select relevant context. The aggregated context information for segment v_i^{l-1} is computed as:

$$m_i = \sum_j^{T_l} \alpha_{ij}^{gtci} \cdot \tilde{c}_j^l, \quad \alpha_{ij}^{gtci} = \frac{exp({v_i^{l-1}}^T \cdot \tilde{c}_j^l + {s_i}^T \cdot \tilde{c}_j^l)}{\sum_j^{T_l} exp({v_i^{l-1}}^T \cdot \tilde{c}_j^l + {s_i}^T \cdot \tilde{c}_j^l)} \quad (3)$$

where the cross-modal feature s_i helps to reassign the semantic coefficient α_{ij}^{gtci}. Subsequently, we build the context gate g_i that controls the flow of aggregated context m_i to v_i^{l-1}, and update the fine segment representations:

$$g_i = \sigma(W_3[v_i^{l-1}, m_i] + b_3), \quad v_i^l = (1-g_i) \odot v_i^{l-1} + g_i \odot m_i \quad (4)$$

Consequently, we obtain $V^{(l)}$ that integrate context features $\tilde{C}^{(l)}$. By performing L DCIN layers, we gradually integrate increasing levels of temporal context into fine segment features and learn final context-aware segment features $V^{(L)} = \{v_i^L\}_{i=1}^T$.

Context-Sensitive Constraint. Fine-grained emotion localization requires the model to discriminate similar emotions such as embarrassment and disquietment. For human beings, they will turn to the rich context that implicates multi-granularity event semantics to disambiguate the emotions. Motivated by this insight, we propose a novel context-sensitive constraint to encourage the model to learn more emotion-discriminative context. Intuitively, we can use the original segment representations $V^{(0)}$ and the context-aware segment representations $V^{(L)}$ to predict the emotion class probability distributions, respectively. If the model is sensitive to the context, the predicted probability will change greatly after integrating the multi-granularity context. Thus, the distance between the two predicted probability distributions should be far. Specifically, given the original segment representations $V^{(0)} = \{v_i^0\}_{i=1}^T$ and the context-aware segment representations $V^{(L)} = \{v_i^L\}_{i=1}^T$, we compute the probability of each emotion class for v_i^0 and v_i^L as:

$$P(E|v_i^0) = f(W_4 v_i^0 + b_4), \; P(E|v_i^L) = f(W_4 v_i^L + b_4) \quad (5)$$

where $P(E|\cdot) \in R^{N\times 1}$ and N is the number of emotion classes. Next, we adopt Euclidean distance to measure the context sensitivity for a pair of v_i^0 and v_i^L as:

$$d(v_i^0, v_i^L) = ||P(E|v_i^L) - P(E|v_i^0)|| \quad (6)$$

Then the context-sensitive constraint loss is formulated as:

$$\mathcal{L}_{cs} = \sum_i^T max(0, \Delta - d(v_i^0, v_i^L)) \quad (7)$$

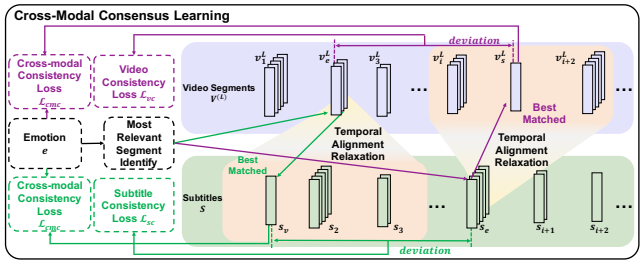

Figure 3: The Cross-Modal Consensus Learning framework.

where Δ is the margin hyper-parameter. By minimizing $\mathcal{L}_{cs}$, we enlarge the distance between the two distributions, encouraging the model to be more sensitive to the context.

3.2 Cross-Modal Consensus Learning

Imagining seeing the subtitle "thanks for your delicate gift", we will infer the visual situation that somebody is happy and vice versa. Inspired by this observation, we propose a cross-modal consensus learning (CCL) paradigm by leveraging the semantic consensus between the aligned video and subtitle. As shown in Figure 3, given the ground-truth emotion label e without temporal annotation, the model first identifies the most relevant video segment. The model then uses its temporally aligned subtitle sentence to predict the most possible emotion. The visual and linguistic modalities are semantically consistent only if the predicted emotion label is the same as the ground-truth.

Specifically, given the ground-truth emotion label e, we first identify the most relevant segment v_e^L as:

$$v_e^L = \underset{v_i^L}{argmax}\, P(e|v_i^L) \quad (8)$$

Then, we retrieve the subtitle sentence s_v that is temporally aligned with v_e (following, we omit the superscript L for simplicity). Next, we compute the emotion class probability distribution based on s_v as $P(E|s_v)$, and maintain the consensus of the score distributions based on v_e and s_v as:

$$\mathcal{L}_{cmc} = -\sum_k^N P(e_k|v_e) log P(e_k|s_v) \quad (9)$$

Where N is the number of emotion classes, and $\mathcal{L}_{cmc}$ is the cross-modal consensus loss. Here we use $P(E|v_e)$ as the pseudo labels, and minimize the cross-entropy loss between them to encourage the semantic consensus. Besides from emotion to video to subtitle, we can also start from emotion to subtitle to video. Let s_e denote the most relevant subtitle sentence and v_s is the aligned video segment. The overall cross-modal consensus loss is:

$$\mathcal{L}_{cmc} = -\sum_k^N [P(e_k|v_e) log P(e_k|s_v) + P(e_k|s_e) log P(e_k|v_s)] \quad (10)$$

Ideally, the most relevant segment v_e should be the same as the segment v_s retrieved from the subtitle side. Thus, we penalize deviation between v_e and v_s, which encourages the semantic consensus on video. The video consensus loss is defined as:

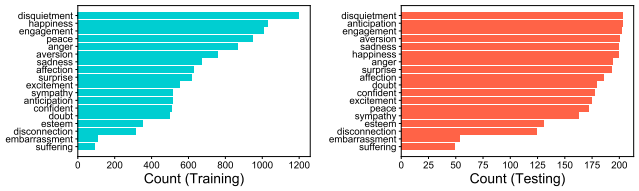

Figure 4: The distribution of 18 emotion classes.

$$\mathcal{L}_{vc} = ||idx(v_e) - idx(v_s)||^2 \tag{11}$$

where $idx(\cdot)$ is the segment index. In a similar manner, we can define the subtitle consensus loss as:

$$\mathcal{L}_{sc} = ||idx(s_e) - idx(s_v)||^2 \tag{12}$$

Temporal Alignment Relaxation. We observe that sometimes people might refer to previous or forthcoming visual events, so the temporal alignment between segment and subtitle may be noisy. In this regard, to alleviate the misalignment noise, we relax the hard temporal alignment constraint and encourage the model to dynamically learn the alignment from our cross-modal consensus learning paradigm. Concretely, for segment v_e, we first retrieve its temporal aligned subtitle sentence s_v. We then take the Q sentences closest to s_v in time as the candidate set $Q(s_v)$. Next, we compute the semantic alignment score between v_e and $s_q \in Q(s_v)$:

$$score(v_e, s_q) = cos(v_e, s_q) - \beta||idx(v_e) - idx(s_q)|| \tag{13}$$

where $cos(\cdot)$ is the cosine similarity and the second term is the index distance. Finally, we select the best match subtitle sentence as s_v^*:

$$s_v^* = \underset{s_q \in Q(s_v)}{argmax}\ score(v_e, s_q) \tag{14}$$

And we can obtain the v_s^* in a similar manner. Finally, We use the s_v^* and v_s^* to compute the losses in Equation 10 - 12.

To differentiate through the cycle, previous methods are usually implemented as soft retrievals (also viewed as attention mechanism). Differently, we implement the *argmax* operation as a "mask" matrix that keeps track of where the maximum of the matrix is. And we empirically observe better performance on the "mask" version.

3.3 Training and Inference

Training. The final training loss for the overall model is:

$$\mathcal{L} = \lambda_1\mathcal{L}_{cs} + \lambda_2\mathcal{L}_{cmc} + \lambda_3\mathcal{L}_{vc} + \lambda_4\mathcal{L}_{sc} \tag{15}$$

Inference. Given the above $V^{(L)}$ and S, the final emotion-segment matching score m_i^e is defined as:

$$m_i^e = \frac{1}{2}(P(e|v_i^L) + P(e|s_i)) \tag{16}$$

where $m_i^e \in R$ is the final matching score between segment i and emotion label e, and $M_i = \{m_i^e\}_{e=1}^N \in R^{N\times 1}$ is the matching score of segment i. We first compute $\{M_i\}_{i=1}^T$ for segment sequence. Then, the emotions where the matching score is above threshold $m_i^e > \gamma_1$ are considered selected. Next, for each selected emotion e^*, we choose the video segment v^* that has the highest matching score with e^*. Finally, we consider the segments adjacent to v^*. If their matching scores are above threshold $m_i^{e^*} > \gamma_2$, we group them iteratively to form the final predicted temporal boundaries for e^*.

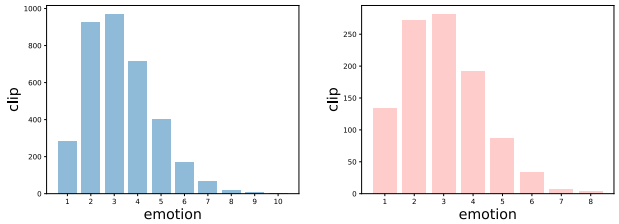

Figure 5: The number of emotions per clip.

4 DATASET

4.1 Fine-Grained Emotion Category Generation.

Existing datasets for emotion recognition are mainly based on still images and classify emotions according to 6 categories. In this work, we introduce a novel task that aims to localize fine-grained emotion in videos. Although several video-based emotion recognition datasets have been proposed, they mainly focus on single-person narratives recorded in controlled lab settings. In contrast, the recently released MovieGraphs dataset [60] contains rich real-world situations, diverse character interactions, and fine-grained emotion annotations. Thus, we evaluate our approach on it. MovieGraphs dataset provides detailed graph-based annotations of social situations for 7637 clips in 51 movies. The dataset was collected and manually annotated using crowd-sourcing methods. The emotion labels are represented as attribute nodes of actors. We extract 239 available emotion labels from all clips and group them into 18 discrete emotion classes. Specifically, we first use word connections (synonyms, relevance, affiliations) and the inter-dependence of a group of words (psychological research and affective computing) [15, 29] to form word-groupings. Then, we perform multiple iterations and cross-referencing with dictionaries and research in affective computing. Finally, we obtain the 18 emotion categories. The details on the definition of each emotion category and the grouped emotions in each category are provided in supplementary materials.

4.2 Dataset Annotation.

We first split and clean the emotion localization samples. As only a few emotion labels in the MovieGraphs dataset have temporal annotations, we develop an annotation tool and ask human annotators to provide the temporal boundaries of emotions in video clips. Before officially annotating, we ask workers to annotate the same set of clips according to provided instructions and examples. For each annotation, we compute its average overlap with the annotations from other workers. If the average overlap is lower than a threshold, we will disregard the annotation. We choose the temporal intersection of consistent annotations as ground-truth. We also manually check the annotations that do not meet the consistency. Overall, 61% of emotions are annotated by at least three workers, and 83% of emotions are annotated by at least two workers. Finally, we take the annotated samples as the testing data.

Method	R@0.5	R@0.7	mAP	mIoU
Random	0.17	0.07	0.19	0.15
Subtitle-Only	5.84	3.10	7.49	5.71
UNets w/o subtitle	1.86	0.50	2.36	1.85
UNets [61]	7.06	2.83	8.38	6.63
3C-Net w/o subtitle	3.69	1.30	6.59	5.09
3C-Net [53]	7.63	2.40	11.42	8.95
ASL w/o subtitle	4.80	1.73	9.25	7.29
ASL [50]	9.56	3.13	14.85	11.35
XML w/o subtitle	7.26	3.07	9.20	7.81
XML [34]	14.30	5.58	19.42	17.27
WSSL w/o subtitle	2.07	0.46	2.52	1.99
WSSL [11]	6.80	2.91	8.82	7.01
DCIN w/o subtitle	13.08	4.51	19.96	16.63
DCIN-CCL	**19.21**	**7.16**	**28.59**	**22.73**

Table 1: Results on the MovieGraph dataset.

4.3 Dataset Statistics

The average duration of video clips is 44.28s and the average number of emotion labels is 3.6 per clip. We re-split the dataset into training (39 movies) and testing (12 movies). The training set consists of 11193 emotions, and the testing set consists of 3003 emotions with temporal annotations. Because some emotion classes are relatively rare in daily life, the distribution of emotion classes is not completely balanced. To comprehensively evaluate the model performance on all classes, we attempt to make the distribution of testing data relatively balanced. The distribution of 18 emotion types is illustrated in Figure 4. We also propose the number of emotions per clip over all training and testing movies in Figure 5.

5 EXPERIMENTS

5.1 Experimental Setup

Implementation Details. For video, we use the ResNeXt-101 model [20] pre-trained on the kinetics-400 dataset as [31]. For subtitle, we employ a pre-trained BERT [10] and perform max pooling over each sentence to get the sentence representations. We set the dimension of segment and subtitle representations to 384. For the visual frames that are not aligned with any subtitles, we assign them to the neighboring segment-subtitle pair. For the hyper-parameters, we set Δ to 0.5, β to 0.1, and set $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ to 0.001, 1.0, 1.0, and 0.7, respectively. During training, we set the batch size to 32 and use Adam as optimizer [12], where the learning rate is set to $1e^{-4}$.
Evaluation Metrics. We employ **R@IoU**, **mIoU**, and **mAP** as evaluation metrics. The **R@IoU** is recall at various thresholds of the temporal Intersection over Union (IoU). The R@IoU measures the percentage of predictions that have IoU with ground-truth larger than the thresholds. Here we set recall to 1, and temporal IoU threshold values to $\{0.5, 0.7\}$. **mAP** is the average precision over various IoU thresholds. **mIoU** is the average IoU between the predicted segments and ground-truth. For **mAP** and **mIoU**, we set temporal IoU threshold values to $\{0.1, 0.3, 0.5, 0.7\}$.
Baselines. We compare the proposed approach with a number of strong baselines from relevant video-and-language tasks. Only publicly available models are used to calculate these metrics. Since the most related task with ours is action localization, we extend the existing weakly-supervised action localization (WSAL) approaches **UntrimmedNets** [61], **3C-Net** [53], and **ASL** [50] as the baselines. Considering these baselines do not take subtitles as input, we implement two versions: 1) ignore subtitles directly; 2) fuse the subtitle features with aligned segment features. **UntrimmedNets** first learns a video-level classification and then selects frames with high classification activation as action locations. **3C-Net** adopts a classification loss to ensure the separability, a center loss to reduce inter-class variations, and a counting loss to delineate adjacent action sequences. **ASL** learns with a class-agnostic task to predict which frames will be selected by the classifier.

	Method	R@0.5	mAP	mIoU
1	Backbone	8.79	12.36	9.66
2	+ Coarse-Fine	10.89	16.39	12.89
3	+ TCDR	11.19	17.64	13.79
4	+ GTCI (w/o CCF)	12.99	20.85	16.62
5	+ CCF = **DCIN**	13.99	22.02	17.36

Table 2: Performance comparison by varying the individual components of the DCIN.

We also extend video-subtitle moment retrieval model **XML** [34] and weakly-supervised temporal grounding model **WSSL** [11] as baselines. **XML** is a recently proposed transformer-based method for TV show retrieval, which first encodes video and subtitle representation separately via two self-encoders, and then builds the cross-modality context representation via two cross-encoders. Here, we use the emotion label to attend to the above fused context features of videos and subtitles. To facilitate weakly-supervised learning, we adopt a multi-instance learning method [9, 76] to train the **XML** model. **WSSL** is a cycle system with a pair of dual problems: event captioning and sentence localization. Here, we train **WSSL** to reconstruct the emotion label as weakly-supervised objective. To show the importance of using both videos and subtitles, we compare baselines with their corresponding video-only variants and extend a standard span-based QA model [25] as **subtitle-only** baseline. For **DCIN w/o subtitle**, we replace the cross-modal consensus learning with the weakly-supervised learning loss from **3C-Net**.

5.2 Results

We compare our approach to the state-of-the-art WSAL methods, video-subtitle moment retrieval, and weakly-supervised temporal sentence grounding. We summarize the results in Table 1. From the results, we can see that our method significantly outperforms the baselines, and the superiority is consistently observed on all metrics. We notice that our **DCIN w/o subtitle** also surpasses baselines on mAP, indicating the effectiveness of our DCIN on multi-granularity temporal context modeling. When using only subtitles to localize emotions, the span-based QA model only achieves 7.49% on mAP. Also, we observe consistent improvement from subtitles on all baselines. These indicate the importance of multi-modal context.

Furthermore, all adapted baselines from three relevant tasks perform poorly on fine-grained emotion localization. We speculate the main reasons are three folds: 1) As the actions and events have more consistent visual patterns, the methods in WSAL and temporal sentence grounding make predictions for each segment separately, ignoring the multi-granularity context. In contrast, we adaptively integrate different granularities of context into segment

	TAR		Loss Terms			R@0.5	mAP	mIoU
	Hard	Relaxed	$\mathcal{L}_{cs}$	$\mathcal{L}_{vc}$	$\mathcal{L}_{sc}$			
1	✓					13.99	22.02	17.36
2		✓				14.99	24.20	19.14
3		✓	✓			16.55	25.71	20.29
4		✓		✓		17.38	26.48	21.04
5		✓			✓	16.25	25.13	19.84
6		✓	✓	✓		18.32	27.40	21.72
7		✓	✓		✓	16.75	26.86	21.20
8		✓		✓	✓	18.38	27.62	21.75
9		✓	✓	✓	✓	**19.21**	**28.59**	**22.73**

Table 3: Ablation studies of the temporal alignment relaxation and the proposed loss terms.

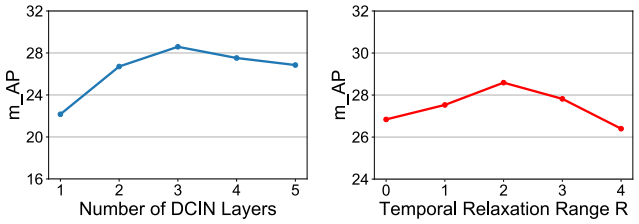

Figure 6: Ablation studies with respect to the number of DCIN layers and the temporal relaxation range.

representations in a hierarchy. 2) The methods in WSAL and temporal sentence grounding are designed for localizing actions or events in pure videos without subtitles. They fail to effectively leverage the rich complementary information in subtitles. Differently, our approach utilizes the cross-modality consensus between video and subtitle to form an effective weakly-supervised learning paradigm. 3) Emotion classes have much higher inter-class similarity than action classes. The methods from all three tasks fail to learn discriminative context, which is crucial for inferring fine-grained information. Instead, we propose a context-sensitive constraint to encourage the model to learn emotion-discriminative context.

5.3 In-Depth Analysis

Effectiveness of Individual Component. We first investigate the contribution of the Dilated Context Integrated Network (DCIN) in Table 2. We start with the backbone model and gradually add the Coarse-Fine architecture, Temporal Context Dependency Reasoning (TCDR), Gated Temporal Context Integration (GTCI), and cross-modal context filtering mechanism (CCF) to form complete DCIN. To clearly distinguish the improvement, we only use the basic cross-modal consensus loss ($\mathcal{L}_{cmc}$) to train these ablation models without temporal relaxation and other proposed losses. As the transformer-based [59] model is a powerful model that effectively captures no-local context, we use it as the backbone for context modeling. It takes video segment sequences $\{v_i\}_{i=1}^{T}$ and subtitle sentence sequences $\{s_i\}_{i=1}^{T}$ as input, performs self-attention and cross-modal attention on them, and finally outputs the context-aware segment and subtitle representations.

As shown in Table 2, the performance increases consistently, indicating the effectiveness of each component. Overall, our DCIN takes up 9.66% of the gain on mAP. Particularly, the results from

Method	THUMOS-14			ActivityNet Captions		
	R@0.3	R@0.5	R@0.7	R@0.3	R@0.5	mIoU
3C-Net [53]	40.9	24.6	7.7	-	-	-
TSCN [64]	47.8	28.7	10.2	-	-	-
ASL [50]	51.8	31.1	11.4	-	-	-
WSSL [11]	-	-	-	41.98	23.34	28.23
SCN [47]	-	-	-	47.23	29.22	-
VGN [76]	-	-	-	50.12	31.07	-
DCIN	50.3	29.8	11.9	46.72	28.19	33.74

Table 4: Transferability of our DCIN.

Row 2 to Row 4 suggest that our DCIN can better model the complex temporal context by adaptively integrating multi-granularity temporal context in a coarse-fine architecture. The results of Row 5 validate the superiority of the cross-modal context filtering mechanism, which utilizes the multi-modality context to better guide the context integration process.

We then verify the strength of our temporal alignment relaxation (TAR) and the proposed losses in Table 3. We start with the complete DCIN (Row 1). The results of Row 2 show that TAR improves the DCIN by dynamically learning the cross-modal alignment, which alleviates the misalignment noise during CCL. Furthermore, the results from Row 3 to Row 5 validate that each loss is helpful for emotion localization. Specifically, context-sensitive constraint ($\mathcal{L}_{cs}$) takes up 6% of the relative gain on mAP and mIoU, the video consistency loss $\mathcal{L}_{vc}$ contributes 1.69% and 1.43% to the improvement on mAP and mIoU, respectively, and the subtitle consistency loss $\mathcal{L}_{sc}$ takes up 1.15% and 0.91% of gain on mAP and mIoU, respectively. In the end, the results from Row 6 to Row 9 suggest that the proposed losses can promote the cross-modal consensus and context sensitivity in a mutually rewarding way.

Impact of DCIN Layer Numbers. Figure 6 presents the results across the number of DCIN layers, where performance increases until 3 layers and decreases afterward. This indicates that 3 DCIN layers can capture enough granularities of temporal context for fine-grained emotion understanding.

Analysis on Temporal Alignment Relaxation. We explore the impact of temporal relaxation range, where we consider the R nearest subtitle (segment) to the central subtitles (segments) in each side as candidates. $R = 0$ corresponds to the hard temporal alignment version. As shown in Figure 6, the performance keeps increasing when the R is increased from 0 to 2. When we continue to increase the R, too many candidates introduce larger noise.

5.4 Transferability to Different Tasks

We further evaluate DCIN on temporal action localization and temporal sentence grounding tasks to illustrate its superiority on multi-granularity temporal context modeling. For temporal action localization, we add fully-connected layers to predict the confidence score of each segment feature encoded by DCIN, and directly adopt the weakly-supervised learning paradigm from 3C-Net to train DCIN. For temporal sentence grounding, we first adopt our DCIN to obtain context-aware segment representations, then generate multiple proposals as 2D-TAN [66], and finally use a multi-instance learning objective to train DCIN.

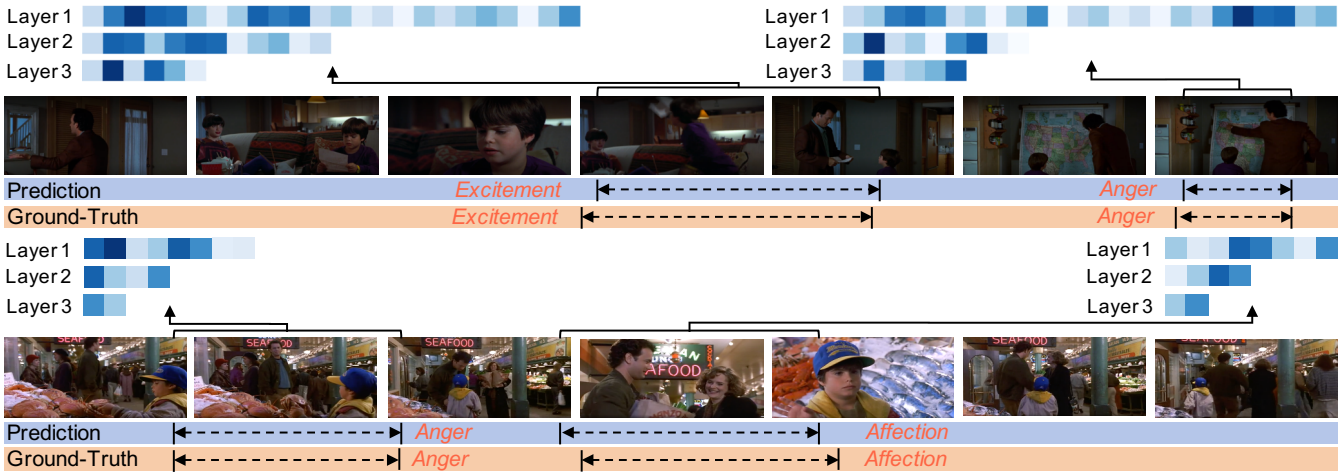

Figure 7: Qualitative examples of our proposed model.

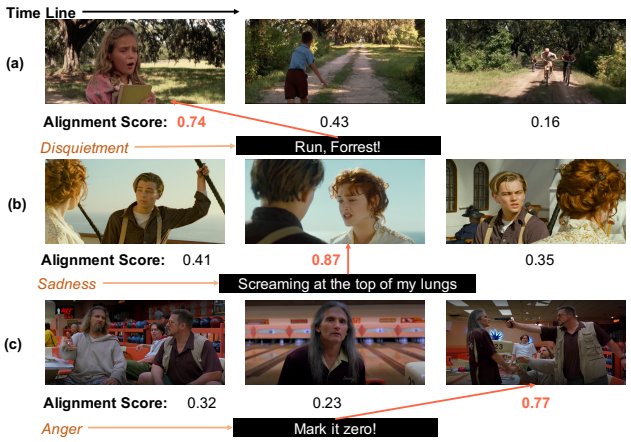

Figure 8: Qualitative examples of CCL.

Table 4 summarizes the temporal action localization performance on THUMOS-14 [17] dataset and the temporal sentence grounding performance on ActivityNet Captions [30] dataset. We notice that the coarse-fine two stream architecture for adaptively multi-granularity temporal dynamics reasoning can also benefit the accurate temporal action localization, improving R@0.7 from 11.4% to 11.9%. Meanwhile, our DCIN can achieve comparable performance on temporal sentence grounding task.

5.5 Qualitative Analysis

Case Study and Visualization of DCIN. Figure 7 visualizes two qualitative examples. Evidently, our model can produce accurate temporal boundaries for the TEL task. For a more intuitive view of how DCIN adaptively integrates multi-granularity context, we also visualize the context gate values of the target segments at different DCIN layers, which reflect the context that the target segments focus on. The context at different time scales is produced by the Coarse stream at different DCIN layers. For instance, for the segment corresponding to emotion *Excitement* (Figure 7.a), its context gates are well activated for the context at all layers that contains the information of the boy happily receiving a letter.

Visualization of CCL. In Figure 8, we show three examples of how the CCL computes the semantic alignment between subtitle and video. For instance, in (a), the subtitle at the second column is the most relevant subtitle for emotion *disquietment* based on $P(e|s_i)$, and the segment at the second column is the temporally aligned segment. The left column and right column correspond to the previous and forthcoming segments, respectively. Below the segments are the semantic alignment scores between them and the subtitle. We can see that the subtitle is said by the girl in the previous segment. If we only follow the hard temporal alignment relationship, we will optimize the segment at the middle column to predict high a score for emotion *disquietment*, which might confuse the model. By considering the temporal alignment relaxation on a neighboring window of it, our CCL adaptively matches the semantic aligned segment (the left column) based on the learned semantic alignment scores.

6 CONCLUSIONS

In this paper, we define a novel task of temporal emotion localization, which fosters deeper investigations in emotion understanding and video-and-language reasoning. To solve the challenges in the task, we propose a novel dilated context integrated network to adaptively integrate multi-granularity temporal context in a hierarchy, as well as a cross-modal consensus learning paradigm for weakly-supervised learning. The experimental results show the effectiveness and transferability of the proposed framework.

ACKNOWLEDGMENT

This work has been supported in part by National Key Research and Development Program of China (2018AAA0101900), Zhejiang NSF (LR21F020004), Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, Key Research and Development Program of Zhejiang Province, China (No. 2021C01013), Chinese Knowledge Center of Engineering Science and Technology (CKCEST). We thank all the reviewers for valuable comments.

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[2] Afsheen Rafaqat Ali, Usman Shahid, Mohsen Ali, and Jeffrey Ho. 2017. High-level concepts for affective understanding of images. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 679–687.

[3] Aron K Barbey, Roberto Colom, and Jordan Grafman. 2014. Distributed neural system for emotional intelligence revealed by lesion mapping. *Social cognitive and affective neuroscience* 9, 3 (2014), 265–272.

[4] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. 2019. End-to-end, single-stream temporal action detection in untrimmed videos. In *Procedings of the British Machine Vision Conference 2017*. British Machine Vision Association.

[5] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. 2017. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2911–2920.

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 961–970.

[7] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[8] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. 2018. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1130–1139.

[9] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. 2020. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308* (2020).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. *Advances in Neural Information Processing Systems* 31 (2018).

[12] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research* 12, 7 (2011).

[13] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 2016. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*. Springer, 768–784.

[14] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. 2016. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5562–5570.

[15] Enrique García Fernández-Abascal, Beatriz García Rodríguez, María Pilar Jiménez Sánchez, María Dolores Martín Díaz, and Francisco Javier Domínguez Sánchez. 2010. *Psicología de la emoción*. Editorial Universitaria Ramón Areces.

[16] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*. 5267–5275.

[17] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. 2015. THUMOS Challenge: Action Recognition with a Large Number of Classes. http://www.thumos.info/.

[18] Jiannan Guo, Yangyang Kang, Yu Duan, Xiaozhong Liu, Siliang Tang, Wenqiao Zhang, Kun Kuang, Changlong Sun, and Fei Wu. 2022. Collaborative Intelligence Orchestration: Inconsistency-Based Fusion of Semi-Supervised Learning and Active Learning. *arXiv preprint arXiv:2206.03288* (2022).

[19] Jiannan Guo, Haochen Shi, Yangyang Kang, Kun Kuang, Siliang Tang, Zhuoren Jiang, Changlong Sun, Fei Wu, and Yueting Zhuang. 2021. Semi-supervised active learning for semi-supervised models: Exploit adversarial examples with graph-based virtual labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2896–2905.

[20] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6546–6555.

[21] Dilana Hazer, Xueyao Ma, Stefanie Rukavina, Sascha Gruss, Steffen Walter, and Harald C Traue. 2015. Emotion elicitation using film clips: Effect of age groups on movie choice and emotion rating. In *International Conference on Human-Computer Interaction*. Springer, 110–116.

[22] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 2016. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1914–1923.

[23] Guillaume Herbet, Gilles Lafargue, François Bonnetblanc, Sylvie Moritz-Gasser, Nicolas Menjot de Champfleur, and Hugues Duffau. 2014. Inferring a dual-stream model of mentalizing from associative white matter fibres disconnection. *Brain* 137, 3 (2014), 944–959.

[24] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[25] Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2017. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. *arXiv preprint arXiv:1711.07341* (2017).

[26] Ziqi Jiang, Shengyu Zhang, Siyuan Yao, Wenqiao Zhang, Sihan Zhang, Juncheng Li, Zhou Zhao, and Fei Wu. 2022. Weakly-supervised Disentanglement Network for Video Fingerspelling Detection. In *ACM MM*.

[27] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[28] Ming Kong, Qing Guo, Shuowen Zhou, Mengze Li, Kun Kuang, Zhengxing Huang, Fei Wu, Xiaohong Chen, and Qiang Zhu. 2022. Attribute-aware interpretation learning for thyroid ultrasound diagnosis. *Artificial Intelligence in Medicine* (2022), 102344.

[29] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2017. EMOTIC: Emotions in Context dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 61–69.

[30] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*. 706–715.

[31] Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. 2020. Learning interactions and relationships between movie characters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9849–9858.

[32] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[33] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696* (2018).

[34] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*. Springer, 447–463.

[35] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. 2022. Fine-Grained Semantically Aligned Vision-Language Pre-Training. *arXiv preprint arXiv:2208.02515* (2022).

[36] Jiacheng Li, Siliang Tang, Juncheng Li, Jun Xiao, Fei Wu, Shiliang Pu, and Yueting Zhuang. 2020. Topic adaptation and prototype encoding for few-shot visual storytelling. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4208–4216.

[37] Juncheng Li, Siliang Tang, Fei Wu, and Yueting Zhuang. 2019. Walking with mind: Mental imagery enhanced embodied qa. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1211–1219.

[38] Juncheng Li, Siliang Tang, Linchao Zhu, Haochen Shi, Xuanwen Huang, Fei Wu, Yi Yang, and Yueting Zhuang. 2021. Adaptive hierarchical graph reasoning with semantic coherence for video-and-language inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1867–1877.

[39] Juncheng Li, Xin Wang, Siliang Tang, Haizhou Shi, Fei Wu, Yueting Zhuang, and William Yang Wang. 2020. Unsupervised reinforcement learning of transferable meta-skills for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12123–12132.

[40] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. 2022. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3032–3041.

[41] Mengze Li, Ming Kong, Kun Kuang, Qiang Zhu, and Fei Wu. 2020. Multi-task attribute-fusion model for fine-grained image recognition. In *Optoelectronic Imaging and Multimedia Technology VII*, Vol. 11550. SPIE, 114–123.

[42] Mengze Li, Kun Kuang, Qiang Zhu, Xiaohong Chen, Qing Guo, and Fei Wu. 2020. IB-M: A Flexible Framework to Align an Interpretable Model and a Black-box Model. In *BIBM*.

[43] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Wenming Tan, Jin Wang, Peng Wang, et al. 2022. End-to-End Modeling via Information Tree for One-Shot Natural Language Spatial Video Grounding. *arXiv preprint arXiv:2203.08013* (2022).

[44] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Wenqiao Zhang, Jiaxu Miao, Shiliang Pu, and Fei Wu. 2022. HERO: HiErarchical spatio-tempoRal reasOning with Contrastive Action Correspondence for End-to-End Video Object Grounding. In *ACM MM*.

[45] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3889–3898.

[46] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. 2018. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 3–19.

[47] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11539–11546.

[48] Yuan Liu, Lin Ma, Yifeng Zhang, Wei Liu, and Shih-Fu Chang. 2019. Multi-granularity generator for temporal action proposal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3604–3613.

[49] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. 2019. Gaussian temporal awareness networks for action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 344–353.

[50] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, and Guangwei Yu. 2021. Weakly Supervised Action Selection Learning in Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7587–7596.

[51] Daniel McDuff, Rana El Kaliouby, Jeffrey F Cohn, and Rosalind W Picard. 2014. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing* 6, 3 (2014), 223–235.

[52] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. 2021. Affect2mm: Affective analysis of multimedia content using emotion causality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5661–5671.

[53] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 2019. 3c-net: Category count and center loss for weakly-supervised action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8679–8687.

[54] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. 2018. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6752–6761.

[55] Stephen Pilli, Manasi Patwardhan, Niranjan Pedanekar, and Shirish Karande. 2020. Predicting sentiments in image advertisements using semantic relations among sentiment labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 408–409.

[56] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and emotion* 24, 7 (2010), 1153–1172.

[57] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4631–4640.

[58] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[60] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. 2018. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8581–8590.

[61] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. 2017. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 4325–4334.

[62] Zijun Wei, Jianming Zhang, Zhe Lin, Joon-Young Lee, Niranjan Balasubramanian, Minh Hoai, and Dimitris Samaras. 2020. Learning visual emotion representations from web data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13106–13115.

[63] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. 2019. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7094–7103.

[64] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. 2020. Two-stream consensus network for weakly-supervised temporal action localization. In *European conference on computer vision*. Springer, 37–54.

[65] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. 2020. DeVLBert: Learning Deconfounded Visio-Linguistic Representations. In *MM '20: The 28th ACM International Conference on Multimedia*. ACM, 4373–4382.

[66] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12870–12877.

[67] Shengyu Zhang, Ziqi Tan, Jin Yu, Zhou Zhao, Kun Kuang, Jie Liu, Jingren Zhou, Hongxia Yang, and Fei Wu. 2020. Poet: Product-oriented Video Captioner for E-commerce. In *MM '20: The 28th ACM International Conference on Multimedia*. ACM, 1292–1301.

[68] Shengyu Zhang, Ziqi Tan, Zhou Zhao, Jin Yu, Kun Kuang, Tan Jiang, Jingren Zhou, Hongxia Yang, and Fei Wu. 2020. Comprehensive Information Integration Modeling Framework for Video Titling. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2744–2754.

[69] Wenqiao Zhang, Jiannan Guo, Mengze Li, Haochen Shi, Shengyu Zhang, Juncheng Li, Siliang Tang, Wu Fei, Tat-Seng Chua, and Yueting Zhuang. 2022. BOSS: Bottom-up Cross-modal Semantic Composition with Hybrid Counterfactual Training for Robust Content-based Image Retrieval. https://doi.org/10.48550/ARXIV.2207.04211

[70] Wenqiao Zhang, Haochen Shi, Jiannan Guo, Shengyu Zhang, Qingpeng Cai, Juncheng Li, Sihui Luo, and Yueting Zhuang. 2022. Magic: Multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3335–3343.

[71] Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. 2021. Consensus graph representation learning for better grounded image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3394–3402.

[72] Wenqiao Zhang, Siliang Tang, Yanpeng Cao, Shiliang Pu, Fei Wu, and Yueting Zhuang. 2019. Frame augmented alternating attention network for video question answering. *IEEE Transactions on Multimedia* 22, 4 (2019), 1032–1041.

[73] Wenqiao Zhang, Siliang Tang, Yanpeng Cao, Jun Xiao, Shiliang Pu, Fei Wu, and Yueting Zhuang. 2020. Photo stream question answer. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3966–3975.

[74] Wenqiao Zhang, Xin Eric Wang, Siliang Tang, Haizhou Shi, Haochen Shi, Jun Xiao, Yueting Zhuang, and William Yang Wang. 2020. Relational graph learning for grounded video description generation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3807–3828.

[75] Wenqiao Zhang, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, and Beng Chin Ooi. 2022. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20666–20676.

[76] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems* 33 (2020), 18123–18134.