

Video Event Extraction with Multi-View Interaction Knowledge Distillation

Kaiwen Wei¹, Runyan Du^{2*}, Li Jin^{2†}, Jian Liu³, Jianhua Yin⁴, Linhao Zhang²,
Jintao Liu², Nayu Liu⁵, Jingyuan Zhang⁶, Zhi Guo²

¹College of Computer Science, Chongqing University, Chongqing, China

²University of Chinese Academy of Sciences, Beijing, China

³Beijing Jiaotong University, Beijing, China

⁴School of Computer Science and Technology, Shandong University, Qingdao, China

⁵School of Computer Science and Technology, Tiangong University, Tianjin, China

⁶Kuaishou Technology Inc., Beijing, China

{weikaiwen1997, durunyan08, jinlimails}@gmail.com

Abstract

Video event extraction (VEE) aims to extract key events and generate the event arguments for their semantic roles from the video. Despite promising results have been achieved by existing methods, they still lack an elaborate learning strategy to adequately consider: (1) inter-object interaction, which reflects the relation between objects; (2) inter-modality interaction, which aligns the features from text and video modality. In this paper, we propose a Multi-view Interaction with knowledge Distillation (MID) framework to solve the above problems with the Knowledge Distillation (KD) mechanism. Specifically, we propose the self-Relational KD (self-RKD) to enhance the inter-object interaction, where the relation between objects is measured by distance metric, and the high-level relational knowledge from the deeper layer is taken as the guidance for boosting the shallow layer in the video encoder. Meanwhile, to improve the inter-modality interaction, the Layer-to-layer KD (LKD) is proposed, which integrates additional cross-modal supervisions (i.e., the results of cross-attention) with the textual supervising signal for training each transformer decoder layer. Extensive experiments show that without any additional parameters, MID achieves the state-of-the-art performance compared to other strong methods in VEE.

Introduction

Video event extraction (VEE) is a challenging multi-modal task in video understanding, which aims to identify the events and generate their arguments in a video. For example, as illustrated in Fig. 1 (a), a VEE system should classify the *drag* event and generate $\langle \text{Arg0}, \text{the horse} \rangle$, $\langle \text{Arg1}, \text{the man wearing the helmet} \rangle$ as its event arguments. VEE could drive many downstream applications, such as video description (Krishna et al. 2017; Xu et al. 2016; Wang et al. 2019), visual content retrieval (Miech et al. 2019), and knowledge graphs (Mahon et al. 2020).

Recently, many methods (Feichtenhofer et al. 2019; Xiao, Tighe, and Modolo 2022; Yang et al. 2022; Xiao et al. 2022)

*These authors contributed equally.

†Corresponding Author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

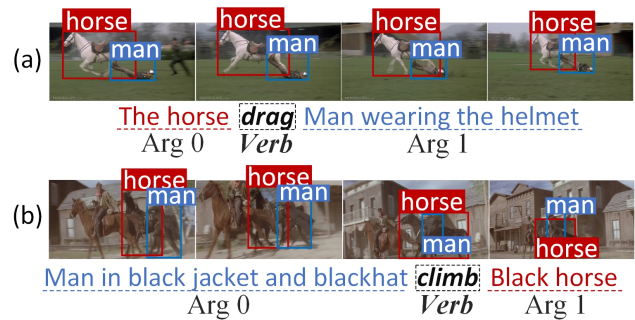


Figure 1: An example of showing the importance of inter-object interaction in VEE. The relative position between “man” and “horse” is similar in the above two scenes, but the two videos show very different events.

have been proposed for the VEE task. For instance, Sadhu et al. (2021) utilized the strong SlowFast model (Feichtenhofer et al. 2019) as the backbone for video feature extraction. Yang et al. (2022) explicitly model the states of objects/entities and their relations in the video. Despite their promising results, existing methods still have shortcomings in considering the interactions from the following views:

(1) **Inter-object interaction:** It reflects the relationship between objects, which is critical to predict verbs and arguments. As shown in Fig. 1, although “horse” and “man” objects exist in both videos, the final event type is totally different because of the way they interact with each other. The existing method (Yang et al. 2022) only models the inter-object interaction with a union calculation following a simple pooling method, which drops most information.

(2) **Inter-modality interaction:** It exists in the the transformer decoder during the decoding process for generating the event arguments, where the output from each transformer layer reflect the interaction between text and video modalities.

Existing methods (Sadhu et al. 2021; Yang et al. 2022) only take the desired output text as the ground truth label for training, and implicitly optimize the cross-modal align-

ment. However, this training strategy neglects that the cross-modal information included in the output features could be served as extra supervision to improve the capability of the inter-modality interaction. To alleviate the above problems, an intuitive way is to annotate a large amount of fine-grained data based on the provided dataset and give the ground-truth for measuring the inter-object and inter-modality interaction. But this approach is time-consuming and laborious.

Based on the above observations, we introduce the Multi-view Interaction with knowledge Distillation (MID) framework to enhance the learning process on the above two interactions. First, to strengthen the inter-object interaction, inspired by previous research (Cornia et al. 2020) that the deeper layer of the encoder has high-level features containing more semantic knowledge than those features in lower layers, we propose a self-Relational Knowledge Distillation (self-RKD) mechanism. It takes the high-level inter-object interaction of the deeper layer to supervise the low-level one in the video encoder by itself. Specifically, self-RKD first measures the degree of interactions between different objects with three kinds of metrics. Then, the computed metrics will be further taken as the distilling signal to transfer the high-level relational knowledge from the deeper layer to the shallow layer. In this manner, the capability of the whole encoder can be promoted by boosting the inter-object interaction layer by layer.

Secondly, to strengthen the inter-modality interaction, we propose a training strategy named Layer-to-layer Knowledge Distillation (LKD). It considers not only single-modal supervisions (textual ground-truth) but also supervisions from cross-modal information, which is included in the output contexts derived from the cross-attention of each transformer decoder layer. Specifically, we first train a teacher model following the regular training procedure with only the ground-truth text as the supervision. Then, the cross-modal contexts derived from cross-attention in each layer of the teacher model are used as additional supervision to train a student model from scratch. The student model who has the same model architecture as the teacher model is randomly initialized and retrained with the combination of textual and cross-modal supervisions. By distilling knowledge from each layer of the trained teacher model to the corresponding layer of the student model, the student model could capture the cross-modal knowledge to directly optimize the inter-modality interactions.

We conduct extensive experiments on the large-scale VidSitu (Sadhu et al. 2021) dataset, and the experimental results have justified the effectiveness of the proposed MID. Particularly, benefiting from the inter-object interaction, MID achieves 2.94% of F1@5 improvements on the test set of the verb classification task compared to existing strong baselines. In addition, the semantic role prediction task enjoys a further boost due to the inter-modality interaction, i.e., 1.32% absolute gains on CIDEr. Additionally, we show that the whole training process does not introduce any additional parameters compared with previous methods. The contributions of this work could be summarized as follows:

- We propose a unified KD-based framework named MID to improve the inter-object and inter-modality interac-

tions. We quantify their particular impacts on the learning for the VEE task.

- This is the first work to introduce the self-Relational KD (self-RKD) to enhance the inter-object interaction. And we design the Layer-to-layer KD (LKD) to enhance the inter-modality interaction.
- Without any additional parameters, the proposed MID achieves the SoTA performance on all the sub-steps of the VEE task. The relevant code will be released to facilitate research in the related area.

Problem Formulation

Given a video clip, a VEE system is required to output a series of events $\{E\}$ contained in a set of frames $\{f_i\}_{i=0}^S$ sampled from the video. Each event E consists of multiple pre-defined roles:

$$E = \{v, \langle r^0, a^0 \rangle, \langle r^1, a^1 \rangle, \dots\}, \quad (1)$$

where v is the verb, r^* are the argument roles, and a^* are the contents of the event arguments.

Extracting each E in VEE is typically modelled as a two-stage pipeline task, consisting of verb classification and semantic role prediction. The former is to predict the verb v from a pre-defined verb set containing N categories according to the video clip. The latter is to generate the event argument a^* with the argument role r^* . For example, the event in Fig. 1 (a) is $v=drag$, and the main event arguments are $\langle \text{AGENT}(\text{ARG}_0), the\ horse \rangle$, $\langle \text{TARGET}(\text{ARG}_1), the\ man\ wearing\ the\ helmet \rangle$, etc.

As shown in Fig. 2, both tasks in the pipeline share the same video encoder, where the SlowFast model (Feichtenhofer et al. 2019) is utilized to generate the grid-like features for each frame. Specifically, the raw video clip will be split into two flows of frames with different sample rates. They pass through their corresponding pathway (Slow or Fast pathway) to generate two kinds of grid-like features \mathbf{g}^{slow} and \mathbf{g}^{fast} with individual convolution blocks. Then, those features are merged together with lateral connection (Feichtenhofer et al. 2019) to form the final grid-like features $\mathbf{g} \in \mathbb{R}^{F \times W \times H \times d}$, where F is the number of the final sampled frames after lateral connection, d is the feature dimension, W and H are the weight and height of the grid feature. Besides, we utilize the object tracking model VidVRD (Gao, Chen, and Huang 2021) to extract the objects' positions, which are represented with the coordinates of their bounding boxes $\mathbf{b} \in \mathbb{R}^{F \times O \times 4}$, where O is the number of detected objects. Finally, the grid-like features \mathbf{g} are further processed by pooling the pixels within the normalized bounding boxes to form the objects' features $\mathbf{p} \in \mathbb{R}^{F \times O \times d}$.

With the objects' features \mathbf{p} and the grid-like feature \mathbf{g} , following Yang et al. (2022), the event-aware video embedding \mathbf{e} is formed through an Embedding Block (EB):

$$\mathbf{e} = \text{EB}(\mathbf{g}, \mathbf{p}). \quad (2)$$

Combining with different decoder modules, the event-aware video embedding \mathbf{e} could be leveraged for verb classification or semantic role prediction tasks.

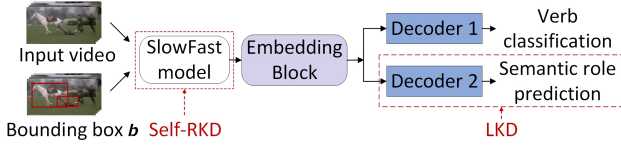


Figure 2: The overall framework of the MID framework.

Method

In this section, we first introduce the sub-tasks of video event extraction, and then present how MID increases: (1) the inter-object interactions via self-RKD, and (2) the inter-modality interactions via layer-to-layer KD.

Verb Classification

The verb classification task takes a 2-second video clip as an input to predict a verb from the pre-defined set. With the encoder to provide event-aware visual embedding \mathbf{e} in Eq. 2, a linear function as the decoder is used to generate a prediction $\mathbf{y}_v \in \mathbb{R}^N$. The cross-entropy loss between $\mathbf{y}_v \in \mathbb{R}^N$ and the ground-truth label $\hat{\mathbf{y}}_v \in \mathbb{R}^N$ is computed:

$$L_v = -\hat{\mathbf{y}}_v \log \mathbf{y}_v, \quad (3)$$

$$\mathbf{y}_v = \text{Softmax}(FC(\mathbf{e})), \quad (4)$$

where L_v is the supervised loss of verb classification, FC is the linear function. In the existing methods, generating event-aware visual embedding \mathbf{e} lacks a comprehensive strategy to activate the inter-object interaction. We present self-RKD to fill this vacancy.

Self-RKD. The previous method (Yang et al. 2022) uses the union box of two objects to gather the pixels within the box and performs average pooling over the pixels to obtain object interaction embedding. The inter-object interaction is modelled by a pooling operator, which inevitably loses useful information and affects the modal performance.

Instead of using a simple pooling strategy, our self-RKD method aims to directly model the interactions between different objects, as shown in Fig. 3. For approaching such purpose, self-RKD explicitly measures the inter-object interaction with a metric function $\varphi(\S, \dagger)$ between two objects and take the higher-level inter-object interaction to boost the low-level one within the SlowFast backbone.

Specifically, we leverage kernel functions to model the interactions between different objects. Three kinds of kernel functions $\varphi(\S, \dagger)$ are provided:

(1) Naïve MMD (Gretton et al. 2008), which measures the distance between two distributions with kernel functions. It reflects the distance between mean embeddings:

$$\varphi(\S, \dagger) = \left| \frac{1}{d} \sum_{k=1}^d \S - \frac{1}{d} \sum_{k=1}^d \dagger \right|, \quad (5)$$

where d is the number of feature dimension of the object.

(2) Dot Production (Lin, RoyChowdhury, and Maji 2015), which calculates the element-wise dot-product of different objects:

$$\varphi(\S, \dagger) = -\S \cdot \dagger^T. \quad (6)$$

(3) Gaussian RBF (Peng et al. 2019), which is a commonly used kernel function whose value depends only on the Euclidean distance from the original space:

$$\varphi(\S, \dagger) = \exp\left(-\frac{\|\S - \dagger\|_2^2}{2\delta^2}\right). \quad (7)$$

Then, for implementing the inter-object interaction knowledge transferring, we indicate \mathbf{g}^l as the grid-like feature output from the l -th layer of the SlowFast model. The coordinates of each object's bounding box extracted from the pre-trained VidVRD (Gao, Chen, and Huang 2021) will be scaled to the space of grid feature \mathbf{g}^l as the normalized bounding box. The ROI align operation (He et al. 2017) is used to obtain the d -dimensional objects' feature from different shape of bounding boxes. The feature of the j -th object generated from the l -th SlowFast layer is $\{\mathbf{p}_{ji}^l\}_{i=0}^F$. With the objects' features, the inter-object interaction \mathbf{I}_o for the l -th SlowFast layer is $\mathbf{I}_o^l \in \mathbb{R}^{F \times O \times O}$, whose element is $\varphi(\mathbf{p}_{j_1 i}^l, \mathbf{p}_{j_2 i}^l)$. O is the number of detected objects, $j_1, j_2 \in [0, O]$ are indexes of objects:

$$\mathbf{I}_o^l = \left\{ \left[\begin{array}{ccc} \varphi(\mathbf{p}_{0i}^l, \mathbf{p}_{0i}^l) & \cdots & \varphi(\mathbf{p}_{0i}^l, \mathbf{p}_{O_i}^l) \\ \vdots & & \vdots \\ \varphi(\mathbf{p}_{O_i}^l, \mathbf{p}_{0i}^l) & \cdots & \varphi(\mathbf{p}_{O_i}^l, \mathbf{p}_{O_i}^l) \end{array} \right] \right\}_{i=1}^F. \quad (8)$$

So far, each layer has its own inter-object interaction \mathbf{I}_o^l . For the l -th layer, its higher-level \mathbf{I}_o^{l+1} will have richer inter-object knowledge than current-level \mathbf{I}_o^l . As a result, we introduce self-RKD to transfer the inter-object knowledge from the higher-level layer to the lower-level layer. The loss of self-RKD in the l -th layer is calculated as follows:

$$L_{self-RKD}^l = \text{KL}(\mathbf{I}_o^{l+1}, \mathbf{I}_o^l), \quad (9)$$

$$L_2^l = \frac{1}{OF} \sum_{i=1}^F \sum_{j=1}^O \|\mathbf{p}_{ji}^{l+1} - \mathbf{p}_{ji}^l\|_2^2, \quad (10)$$

where $\text{KL}(\cdot)$ is the Kullback–Leibler divergence and L_2^l is the L2 loss between the objects' features in the adjacent layers. The L2 loss will provide the value stability of \mathbf{p}_{ji} in training. The overall loss function in the verb classification task is:

$$L_{vb} = \alpha_0 * L_v + \frac{1}{N^v} \sum_{l=1}^{N^v} (\alpha_1 * L_{self-RKD}^l + \alpha_2 * L_2^l), \quad (11)$$

where α_* are the hyper-parameters to make a trade-off among three losses. N^v is the number of SlowFast layers.

Semantic Role Prediction

After training with the verb classification task, the parameters of the video encoder are frozen when training with the semantic role prediction task. The frozen video encoder provides the event-aware video embedding \mathbf{e} . Through a transformer decoder, the event arguments and their argument roles are generated in an auto-regressive manner:

$$\mathbf{y}_s = \text{TransDec}(\mathbf{e}, v, r^0, a^0, \dots), \quad (12)$$

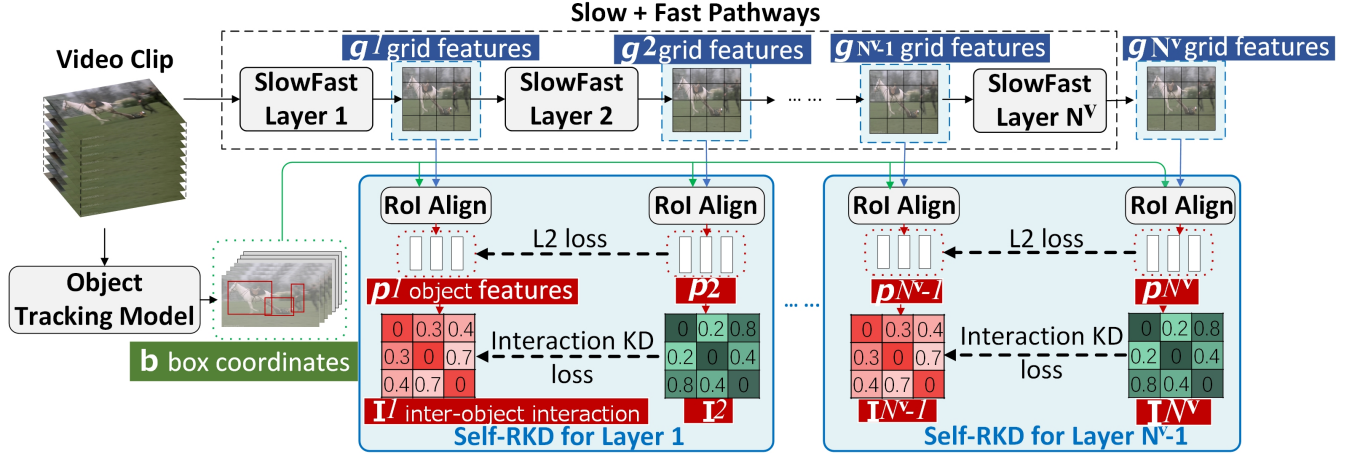


Figure 3: The architecture of the SlowFast model with self-RKD. Between two adjacent SlowFast layers, the L2 loss and interaction KD loss are measured over the objects’ features and inter-object interactions of two layers, respectively.

where $\text{TransDec}(\cdot)$ is the transformer decoder, v is the verb, r^* are the special tokens for each argument role, and a^* are event arguments. $\mathbf{y}_s \in \mathbb{R}^{N_p \times d_{vocab}}$ is the likelihood of the predicted texts. N_p is the length of texts. d_{vocab} is the size of the vocabulary.

The \mathbf{y}_s is generated through multiple stacked transformer decoder layers following a linear function, which projects features of argument sequence \mathcal{P}^{N^l} to their corresponding text, where N^l is the last transformer decoder layer. Specifically, argument sequence \mathcal{P}^{N^l} is the output of the final TransDec layer. In the l -th layer of TransDec , the argument sequence of this layer is $\mathcal{P}^l = \{\mathbf{v}^l, \mathbf{w}_1^l, \dots, \mathbf{w}_{N_p}^l\}$, where \mathbf{v}^l is the feature of verb, and \mathbf{w}_*^l are the features of special tokens (e.g., “[Arg0]”, representing the argument role) or words (e.g., “white horse”, describing the event argument) in the l -th decoder layer.

With the likelihood of the predicted logits \mathbf{y}_s , a cross-entropy loss L_s is calculated between \mathbf{y}_s and the ground-truth words $\hat{\mathbf{y}}_s$ for the semantic role prediction task:

$$L_s = -\frac{1}{N_p} \sum_{i=1}^{N_p} \hat{\mathbf{y}}_s \log \mathbf{y}_s. \quad (13)$$

The general loss L_s of semantic role prediction explicitly supervises the logit of each predicted word. However, the inter-modality interaction in each decoder layer lacks a direct supervising signal for optimization. To a certain extent, the capacity of inter-modality interaction is not fully activated. As a result, we propose the layer-to-layer KD (LKD) to provide an additional soft label from the cross-modal context to make the inter-modality interaction of each layer obtain guidance from a teacher model.

Layer-to-layer KD. LKD is conducted on the TransDec in Eq. 12. The principle of LKD to improve the model performance is similar to transfer learning (Chen et al. 2021b). It provides a shortcut to let the inter-modality interaction of the student quickly adapts to the semantic role prediction task.

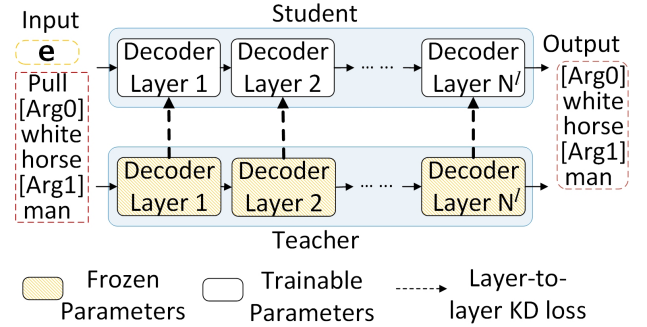


Figure 4: The architecture of the layer-to-layer KD, which provides inter-modality interaction guidance from the teacher to the student at each transformer decoder layer.

Specifically, we first train a teacher model with the regular training process and freeze its parameters. Then we leverage the ground-truth text and the output cross-modal context from each layer of the teacher model to train a randomly initialized student model, which has the same model structure as the teacher model.

In LKD, the inter-modality interaction of student and teacher in the l -th decoder layer is noted as $\mathbf{I}_{m-student}^l$ and $\mathbf{I}_{m-teacher}^l$, respectively. They are the intermediate output of the cross attention at each decoder layer. Essentially, both of them represent the similarity between the event-aware video embedding \mathbf{e} and argument sequence \mathcal{P}^l of the l -th decoder layer. They contain the knowledge about the alignment between visual and textual modalities.

For borrowing such interaction knowledge from the teacher, as illustrated in Fig. 4, we calculate the LKD loss based on two interactions $\mathbf{I}_{m-student}^l$ and $\mathbf{I}_{m-teacher}^l$. Specifically, we use the average pooling method to get the features along the head dimension. Assuming the original multi-head output of the cross-attention is $\mathbf{o} \in$

	Train	Valid	Test-Verb	Test-Role
Clip	118,130	6,630	6,765	7,990
Verb	118,130	66,300	67,650	79,900
Role	118,130	19,890	20,295	23,970

Table 1: The statistic of the dataset.

$\mathbb{R}^{h \times (N_p+1) \times d}$, the output value will be $\hat{o} \in \mathbb{R}^{(N_p+1) \times d}$. Then, the LKD loss for the l -th decoder layer between teacher and student is calculated with their output features $\hat{o}_{student}^l$ and $\hat{o}_{teacher}^l$:

$$L_m^l = \|\hat{o}_{teacher}^l - \hat{o}_{student}^l\|_2^2. \quad (14)$$

With LKD for each decoder layer, we could get an additional supervising signal for optimizing the inter-modality interaction. Combined with the general loss in Eq. 13, the overall loss function can provide better convergence and enhance inter-modality interactions in each decoder layer, which could be formulated as:

$$L_{sem} = \beta_0 * L_s + \beta_1 * \frac{1}{N^l} \sum_{l=1}^{N^l} L_m^l, \quad (15)$$

where β_* are the hyper-parameters to balance the above two losses. N^l is the number of decoder layers.

Experiment

Dataset and Evaluations

We evaluate the models on the VidSitu (Sadhu et al. 2021) dataset, which is a large-scale video understanding dataset with over 130,000 video clips. The event ontology comprises 2,154 verb-senses, and each verb is associated with a minimum of 3 semantic roles. The specific dataset statistics are illustrated in Table 1. The results on the test set are hidden and displayed in the leaderboard¹. As for evaluation metrics, following Yang et al. (2022), for the verb classification task, we calculate the ranking-based metrics including the Accuracy@1, Accuracy@5, Recall@5, and F1@5. Meanwhile, for the semantic role prediction task, CIDEr score (Vedantam, Zitnick, and Parikh 2015) and ROUGE-L (Lin 2004) are computed between predictions and ground-truth descriptions. We compute the micro-averaged CIDEr score for each individual prediction, and we also calculate the macro-averaged CIDEr score over every verb-sense (CIDEr-Verb) and argument-type (CIDEr-Arg).

Implementation

For the verb classification task, we leverage SlowFast (Feichtenhofer et al. 2019) as the feature extractor and follow its frame sampling rate and dimensions of grid features. The batch size is set as 16. We leverage the Adam optimizer with 1e-4 learning rate. The maximum number of objects in each frame is 8. When distilling between different blocks,

¹<https://leaderboard.allenai.org/vidsitu-verbs/submissions/public>

the α_* and β_* in both losses are all set as 1.0. All the experiments are conducted on 4 V100 GPUs. The models are trained for 10 epochs and reported with highest validation F1@5 score. For the semantic role prediction task, the visual event embedding for each video clip remains fixed, and we only train the sequence-to-sequence model. The number of transformer decoder layers is 3. We train the model for 10 epochs and report its performance using the highest validation CIDEr obtained. The optimal hyper-parameters are obtained by grid search. Following Yang et al. (2022), we run the models 10 times with random seeds as 17, 33, 66, 74, 98, 137, 265, 314, 590, 788 due to the high variance of free-form generated argument names.

Baselines

State-of-the-art Models. We compare with the recent state-of-the-art (SoTA) models: (1) **TimeFormer** (Bertasius, Wang, and Torresani 2021), which enables spatiotemporal feature learning directly from a sequence of frame-level patches; (2) **I3D** (Carreira and Zisserman 2017), where filters and pooling kernels of very deep image classification ConvNets are expanded into 3D; (3) **SlowFast** (Feichtenhofer et al. 2019), which contains a slow pathway and a fast pathway for video recognition; For the I3D and SlowFast baselines, we consider the variant with Non-Local blocks (Wang et al. 2018) for comparison. We also compare with three variances of the OSE model (Yang et al. 2022): (4) **OSE-pixel + OME**, which uses object state embedding to track pixel changes; (5) **OSE-pixel/disp + OME**, which uses object state embedding to track both pixel changes and displacements; (6) **OSE-pixel/disp + OME + OIE**, which uses object state embedding and object interaction embedding. For the semantic role prediction task, we also include: (7) **GPT2** (Radford et al. 2019), which is a text-only decoder and has shown potential in many generation tasks. (8) **Video-LLaMA** (Zhang, Li, and Bing 2023), which is a multi-modal framework that empowers large language models for understanding the visual content in the video. Due to the limitation of computational resources, we perform zero-shot inference on Video-LLaMA. The input template of Video-LLaMA is as follows:

Please parse the content of this video. Please generate five categories of event argument “Arg0”, “Arg1”, “Arg2”, “ALoc”, “AScn” for the whole video. Those categories represent “Arg0”: Agent, object performing the action; “Arg1”: Patient, object on which action is performed; “Arg2”: Instrument, Benefactive, Attribute; “ALoc”: location; “AScn”: where the event takes place. Please return the results generated for each category in json format as: {“the event arguments of video”:{“Arg0”:[], “Arg1”:[], “Arg2”:[], “ALoc”:[], “AScn”:[]}}.

Ablation Models. We provide four variances of MID according to the self-RKD methods to model the object-object interactions: (8) **MID (DOT)**, which leverages dot production; (9) **MID (RBF)**, which leverages Gaussian RBF; (10) **MID (MMD)**, which leverage naive MMD; (11) **MID (None)**, which does not utilize any self-RKD method but model the inter-modality interactions with layer-to-layer

Model	Val				Test			
	Acc@1	Acc@5	Rec@5	F1@5	Acc@1	Acc@5	Rec@5	F1@5
TimeSformer	45.91	79.97	23.61	36.46	-	-	-	-
I3D♠	30.17	66.83	4.88	9.10	31.43	67.70	5.02	9.35
SlowFast♠	32.64	69.22	6.11	11.23	33.94	70.54	6.56	12.00
I3D	29.65	60.77	18.21	28.02	29.87	59.10	19.54	29.37
SlowFast	46.79	75.90	23.38	35.75	46.37	75.28	25.78	38.41
OSE-pixel + OME	52.75	83.88	28.44	42.48	52.14	83.84	30.66	44.90
OSE-pixel/disp + OME	53.32	84.00	28.61	42.68	51.88	83.55	30.83	45.04
OSE-pixel/disp + OME + OIE	53.36	83.94	28.72	42.80	52.39	83.47	30.74	44.93
MID (DOT)	52.11	83.33	29.50	43.57	52.31	83.27	31.86	45.90
MID (RBF)	53.14	84.43	30.30	44.60	52.87	83.89	33.01	47.38
MID (MMD)	54.84	85.14	30.68	45.11	53.11	84.55	33.49	47.98

Table 2: Experiment results on verb classification. ♠ indicates those methods does not pre-train on Kinetics-400 dataset.

Model	CIDEr		CIDEr-Verb		CIDEr-Arg		ROUGE-L	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std
GPT2♣	34.67		42.97		34.45		40.08	
I3D♣	47.06		51.67		42.76		42.41	
SlowFast♣	45.52		55.47		42.82		42.66	
Video-LLaMA	33.01		38.14		32.26		37.02	
SlowFast	44.49 ± 2.30		51.73 ± 2.70		40.93 ± 2.42		40.83 ± 1.27	
OSE-pixel + OME	47.82 ± 2.12		54.51 ± 3.00		44.32 ± 2.45		40.91 ± 1.32	
OSE-pixel/disp + OME	48.46 ± 1.84		56.04 ± 2.12		44.60 ± 2.33		41.89 ± 1.12	
OSE-pixel/disp + OME + OIE	47.16 ± 1.71		53.96 ± 1.32		42.78 ± 2.74		40.86 ± 2.54	
MID (None)	48.60 ± 1.94		56.66 ± 1.92		44.98 ± 1.76		41.91 ± 1.98	
MID (DOT)	48.71 ± 2.03		56.84 ± 1.73		45.13 ± 2.99		42.07 ± 1.83	
MID (RBF)	49.12 ± 1.87		57.28 ± 1.94		45.71 ± 2.85		42.24 ± 1.52	
MID (MMD)	49.78 ± 1.91		57.86 ± 2.28		46.49 ± 2.84		42.56 ± 1.13	

Table 3: Experiment results on semantic role prediction. We report the average (Avg) results over 10 runs with standard deviation (Std). The results with ♣ are the single-run performance reported in the VidSitu paper (Sadhu et al. 2021).

KD, utilizing the fixed visual event embedding of OSE-pixel/disp + OME for the semantic role prediction task.

Main Experiment

Verb Classification. The experiment results are shown in Table 2, where we could find that: (1) Both OSE and the proposed MID leverage bounding boxes to track the object’s visual states. They both provide performance gains over those baselines, demonstrating the effectiveness of modelling objects at a finer granularity. (2) Compared to the state-of-the-art method OSE-pixel/disp + OME + OIE, all three variants of MID have achieved a certain level of performance gains in F1@5. Meanwhile, the methods utilizing self-RKD generally have a higher Recall@5, indicating the inter-object interactions could effectively help the models to recall more correct results from the verb candidate set.

Semantic Role Prediction. The experiment results on the validation dataset are illustrated in Table 3, where we could find: (1) OSE-pixel/disp + OME and MID (None) leverage the same visual event embedding generated from the verb classification stage. However, MID (None) outperforms the strong baseline in CIDEr. It illustrates that the LKD

	CIDEr	CIDEr-Verb	CIDEr-Arg	ROUGE-L
MID (None)*	48.46	56.04	44.60	41.89
MID (DOT)*	48.52	56.26	44.81	41.94
MID (RBF)*	48.72	56.52	45.12	41.98
MID (MMD)*	49.13	56.92	46.03	42.11
MID (MMD)	49.78	57.86	46.49	42.56

Table 4: The average experiment results over 10 runs on semantic role prediction. * indicates without LKD.

could effectively model the inter-modality interactions, leading to performance improvements. In addition, the performance further improves when leveraging features from the verb classification task trained with different self-RKD mechanisms, demonstrating that inter-object interaction can also promote argument generation. (2) The large language model-based Video-LLaMA dose not perform well, an important reason is that it may generates long descriptive texts, containing a number of arguments that do not belong to the desired event category. (3) MID (DOT) does not perform rel-

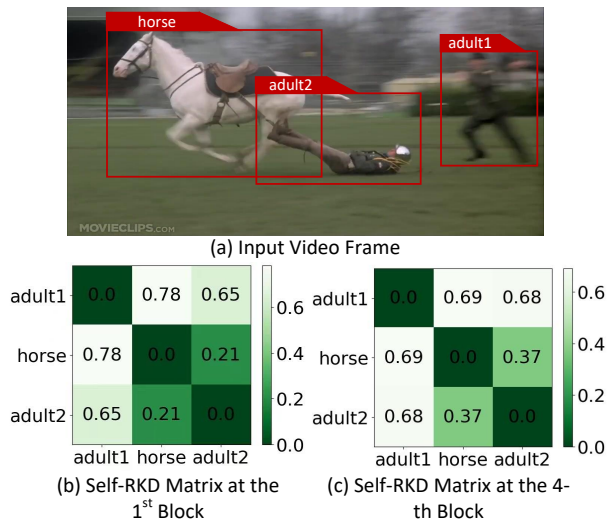


Figure 5: The inter-object interaction visualization.

atively well on both sub-steps of VEE. A possible reason is that using dot production may not be able to precisely model the relationship between objects, thus affecting performance on subsequent tasks. (4) Compared with those strong baselines (such as OSE), which introduces trainable object embeddings, MID does not require any additional parameters, and it brings performance improvement only through KD.

Ablation Study without LKD. We conduct the ablation study by removing the layer-to-layer knowledge distillation (LKD). The experiment results are illustrated in Table 4, where we could find that: compared to those methods that utilise the LKD in Table 3, the model performance meets different degrees of degradation when we remove the LKD, no matter using self-RKD or not. This is consistent with our intuition that adding supervisory information directly to each decoder layer of the transformer would benefit the inter-modality interactions.

Object-object Interaction Visualization. We conduct the experiment to visualize the interactions between different objects, where the bounding boxes and object labels are obtained from the object tracking model VidVRD (Gao, Chen, and Huang 2021). The similarities between object pairs are measured by the MMD metric. As illustrated in Fig. 5 (a), there are three objects in the input video frame: the horse, the running adult1, and the horse-drawn adult2. From Fig. 5 (b)(c), we could find that: since the horse drags the adult2, there is a relatively strong inter-object interaction between horse and adult2. Meanwhile, the similarity between horse and adult2 becomes stronger as the number of model layers deepens. This finding is consistent with the finding in Cornia et al. (2020) that the deeper layer of the feature extractor has high-level features. From this visualization result, we could find that the proposed self-RKD mechanism could effectively capture the inter-object interactions.

Case Study. We list several typical cases in Fig. 6 to demonstrate that the proposed MID enhances the inter-

object and inter-modality interactions. In the first example, there are two people pointing at each other and posing. The proposed MID correctly determines *point* and *gesture* because it models the relationship between the two people in a more appropriate way. Whereas the SoTA model OSE incorrectly classifies the verb category as *speak*. In the second example, there is a couple hugging and kissing each other. Our method successfully captures this inter-object interaction and correctly predicts *embrace* and *hug*. However, OSE only focuses on that they are talking, leading to several errors. In the third example, there is a woman smoking. Compared to the results generated by OSE, MID is able to generate the details of the woman’s clothes. An important factor is that MID effectively models the inter-object and inter-modality interactions, thus more detailed results can be generated. We also compared with the strong video large language model Video-LLaMA (Zhang, Li, and Bing 2023), and we could find that Video-LLaMA may generate event arguments that do not belong to the specific argument category (e.g., the Arg0 type), which leads to difference between the generated results and the ground truth values.

Related Work

Video Event Extraction

Event extraction is a crucial sub-task of information extraction. It is first proposed in the text field (Liu et al. 2023b; Wei et al. 2021; Liu et al. 2023a, 2020), and then expand into image/video (Yatskar, Zettlemoyer, and Farhadi 2016; Pratt et al. 2020; Sadhu et al. 2021), or multimedia (Li et al. 2020; Chen et al. 2021a; Li et al. 2022) fields. Specifically, in the video event extraction domain, Sadhu et al. (2021) introduced a VidSitu dataset for video understanding. A series of follow-up efforts (Xiao, Tighe, and Modolo 2022; Yang et al. 2022; Xiao et al. 2022) were proposed to drive the development of related areas. For example, Yang et al. (2022) explicitly model the states of objects/entities and their relationships in the videos. However, they do not adequately model the interaction between objects (only calculate the union of the bounding boxes and then get the pooling results). Meanwhile, leveraging the supervision of interactions between different modalities at each transformer decoder layer to boost performance is rarely considered in existing methods. To solve these problems, in this work, we introduce self-RKD and LKD, respectively.

Knowledge Distillation

Knowledge distillation (KD) is first proposed by Hinton, Vinyals, and Dean (2015). It has been widely used in many modalities, such as text (Liu et al. 2023c; Wei et al. 2022), image (Passalis and Tefas 2018; Zhuang et al. 2018), and video (Bhardwaj, Srinivasan, and Khapra 2019; Zhang et al. 2020). Zhang et al. (2019) introduced the self distillation mechanism, which computes the association between different CNN block feature maps, but they cannot model the interaction between objects. Müller, Kornblith, and Hinton (2019) pointed out that the KD mechanism could play the role of label smoothing. Several works (Chen et al. 2021b; Zhang et al. 2019; Sun et al. 2020) utilize this property and




		
Our model gesture.01, point.02, wave.01	OSE speak.01, wave.01,talk.01	Ground Truth knock.01,point.0 2,gesture.01
		
Our model kiss.01,embrace. 02, hug.01	OSE look.01,speak.0 1, talk.01	Ground Truth embrace.02, hug.01, talk.01
		
Our model Arg0: woman in a black coat	OSE Arg0: woman	Ground Truth Arg0: woman in a black coat
Video-LLaMA		
Arg0: The man is wearing a suit	Arg0: the woman is wearing a black suit jacket	Arg0: a backlit white cloth- covered table

Figure 6: Result comparison between SoTA and our model.

apply KD to the knowledge transfer field. However, leveraging the cross-modal contexts derived from cross-attention in the teacher model as the soft label to supervise the training process of the student model is rarely considered.

Conclusion

In this work, we quantify the inter-object and inter-modality interactions' impacts on the learning for the video event extraction task. To promote the two interactions, we propose a unified model named MID, which consists of the self-relational knowledge distillation (self-RKD) and the layer-to-layer knowledge distillation (LKD), respectively. Experimental results illustrate that without introducing any additional parameters, the proposed MID achieves the SoTA performance on the large-scale VidSitu dataset. Considering the generality of MID, in the future, we plan to apply it to other event-related video tasks, such as action classification, video description, etc.

Acknowledgments

We sincerely thank all the anonymous reviewers. This research was supported by the National Natural Science Foundation of China (Grant No. 62206267, 62172261).

References

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, 813–824. PMLR.

Bhardwaj, S.; Srinivasan, M.; and Khapra, M. M. 2019. Efficient Video Classification Using Fewer Frames. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 354–363. Computer Vision Foundation / IEEE.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 4724–4733. IEEE Computer Society.

Chen, B.; Lin, X.; Thomas, C.; Li, M.; Yoshida, S.; Chum, L.; Ji, H.; and Chang, S. 2021a. Joint Multimedia Event Extraction from Video and Article. In Moens, M.; Huang, X.; Specia, L.; and Yih, S. W., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 74–88. Association for Computational Linguistics.

Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021b. Distilling Knowledge via Knowledge Review. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 5008–5017. Computer Vision Foundation / IEEE.

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-Memory Transformer for Image Captioning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slow-Fast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV*, 6201–6210. IEEE.

Gao, K.; Chen, L.; and Huang, Y. 2021. Video relation detection via tracklet based visual transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4833–4837.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. J. 2008. A Kernel Method for the Two-Sample Problem. *CoRR*, abs/0805.2368.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2980–2988. IEEE Computer Society.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Niebles, J. C. 2017. Dense-Captioning Events in Videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 706–715. IEEE Computer Society.

Li, M.; Xu, R.; Wang, S.; Zhou, L.; Lin, X.; Zhu, C.; Zeng, M.; Ji, H.; and Chang, S. 2022. CLIP-Event: Connecting Text and Images with Event Structures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 16399–16408. IEEE.

Li, M.; Zareian, A.; Zeng, Q.; Whitehead, S.; Lu, D.; Ji, H.; and Chang, S. 2020. Cross-media Structured Common Space for Multimedia Event Extraction. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, 2557–2568. Association for Computational Linguistics.

- Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.
- Lin, T.; RoyChowdhury, A.; and Maji, S. 2015. Bilinear CNN Models for Fine-Grained Visual Recognition. In *2015 IEEE International Conference on Computer Vision, ICCV, 1449–1457*. IEEE Computer Society.
- Liu, J.; Chen, Y.; Liu, K.; Bi, W.; and Liu, X. 2020. Event Extraction as Machine Reading Comprehension. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, 1641–1651. Association for Computational Linguistics.
- Liu, J.; Zhang, Z.; Guo, Z.; Jin, L.; Li, X.; Wei, K.; and Sun, X. 2023a. Emotion-cause pair extraction with bidirectional multi-label sequence tagging. *Applied Intelligence*, 1–16.
- Liu, J.; Zhang, Z.; Guo, Z.; Jin, L.; Li, X.; Wei, K.; and Sun, X. 2023b. KEPT: Knowledge Enhanced Prompt Tuning for event causality identification. *Knowl. Based Syst.*, 259: 110064.
- Liu, J.; Zhang, Z.; Wei, K.; Guo, Z.; Sun, X.; Jin, L.; and Li, X. 2023c. Event Causality Extraction via Implicit Cause-Effect Interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 6792–6804.
- Mahon, L.; Giunchiglia, E.; Li, B.; and Lukasiewicz, T. 2020. Knowledge Graph Extraction from Videos. In Wani, M. A.; Luo, F.; Li, X. A.; Dou, D.; and Bonchi, F., eds., *19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020, Miami, FL, USA, December 14-17, 2020*, 25–32. IEEE.
- Miech, A.; Zhukov, D.; Alayrac, J.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2630–2640. IEEE.
- Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d’Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 4696–4705.
- Passalis, N.; and Tefas, A. 2018. Learning Deep Representations with Probabilistic Knowledge Transfer. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference*, volume 11215 of *Lecture Notes in Computer Science*, 283–299. Springer.
- Peng, B.; Jin, X.; Li, D.; Zhou, S.; Wu, Y.; Liu, J.; Zhang, Z.; and Liu, Y. 2019. Correlation Congruence for Knowledge Distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV, 5006–5015*. IEEE.
- Pratt, S. M.; Yatskar, M.; Weihs, L.; Farhadi, A.; and Kembhavi, A. 2020. Grounded Situation Recognition. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference*, volume 12349 of *Lecture Notes in Computer Science*, 314–332. Springer.
- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.
- Sadhu, A.; Gupta, T.; Yatskar, M.; Nevatia, R.; and Kembhavi, A. 2021. Visual Semantic Role Labeling for Video Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 5589–5600*. Computer Vision Foundation / IEEE.
- Sun, Z.; Yu, H.; Song, X.; Liu, R.; Yang, Y.; and Zhou, D. 2020. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2158–2170*. Association for Computational Linguistics.
- Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDER: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 4566–4575*. IEEE Computer Society.
- Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition, 7794–7803*.
- Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.; and Wang, W. Y. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 4580–4590. IEEE.
- Wei, K.; Sun, X.; Zhang, Z.; Zhang, J.; Guo, Z.; and Jin, L. 2021. Trigger is Not Sufficient: Exploiting Frame-aware Knowledge for Implicit Event Argument Extraction. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, 4672–4682. Association for Computational Linguistics.
- Wei, K.; Zhang, Z.; Jin, L.; Guo, Z.; Li, S.; Wang, W.; and Lv, J. 2022. HEFT: A History-Enhanced Feature Transfer framework for incremental event detection. *Knowl. Based Syst.*, 254: 109601.
- Xiao, F.; Kundu, K.; Tighe, J.; and Modolo, D. 2022. Hierarchical Self-supervised Representation Learning for Movie Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 9717–9726*. IEEE.
- Xiao, F.; Tighe, J.; and Modolo, D. 2022. MaCLR: Motion-Aware Contrastive Learning of Representations for Videos. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference*, volume 13695 of *Lecture Notes in Computer Science*, 353–370. Springer.
- Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision*

and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 5288–5296. IEEE Computer Society.

Yang, G.; Li, M.; Zhang, J.; Lin, X.; Chang, S.; and Ji, H. 2022. Video Event Extraction via Tracking Visual States of Arguments. *CoRR*, abs/2211.01781.

Yatskar, M.; Zettlemoyer, L.; and Farhadi, A. 2016. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 5534–5542. IEEE Computer Society.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858*.

Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; and Ma, K. 2019. Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV*, 3712–3721. IEEE.

Zhang, Z.; Shi, Y.; Yuan, C.; Li, B.; Wang, P.; Hu, W.; and Zha, Z. 2020. Object Relational Graph With Teacher-Recommended Learning for Video Captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 13275–13285. Computer Vision Foundation / IEEE.

Zhuang, B.; Shen, C.; Tan, M.; Liu, L.; and Reid, I. D. 2018. Towards Effective Low-Bitwidth Convolutional Neural Networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 7920–7928. Computer Vision Foundation / IEEE Computer Society.