Knowledge-enhanced Multi-perspective Video Representation Learning for Scene Recognition

Xuzheng Yu^{1,2}, Chen Jiang², Wei Zhang², Tian Gan¹ Linlin Chao², Jianan Zhao², Yuan Cheng², Qingpei Guo², Wei Chu²

¹Shandong University

²Ant Group

Abstract

With the explosive growth of video data in real-world applications, a comprehensive representation of videos becomes increasingly important. In this paper, we address the problem of video scene recognition, whose goal is to learn a high-level video representation to classify scenes in videos. Due to the diversity and complexity of video contents in realistic scenarios, this task remains a challenge. Most existing works identify scenes for videos only from visual or textual information in a temporal perspective, ignoring the valuable information hidden in single frames, while several earlier studies only recognize scenes for separate images in a non-temporal perspective. We argue that these two perspectives are both meaningful for this task and complementary to each other, meanwhile, external introduced knowledge can also promote the comprehension of videos. We propose a novel two-stream framework to model video representations from multiple perspectives, i.e. temporal and non-temporal perspectives, and integrate the two perspectives in an end-to-end manner by self-distillation. Besides, we design a knowledge-enhanced feature fusion and label prediction method that contributes to naturally introducing knowledge into the task of video scene recognition. Experiments conducted on a real-world dataset demonstrate the effectiveness of our proposed method.

1. INTRODUCTION

With the explosive growth of video data in real-world applications, analysis and reasoning on video content become increasingly important. One important branch is scene recognition from videos, which is to identify scene labels based on the content of videos.

Video understanding itself is complex, including the diversity and redundancy of video contents, and the inherent gap among video multi-modal information. Therefore, the first intuitive challenge is how to effectively and



Figure 1. An illustration of the task of scene recognition.

comprehensively understand the contents of videos for this task. Based on a deep insight of the characteristics of video scene recognition, it is notable that there exists multiperspective information, including global vs local, temporal vs non-temporal, visual vs textural, etc. However, the diversity and discreteness of multiple information make it difficult to distinguish between useful and useless information. Meanwhile, we note that knowledge enhancement has been proven effective in many tasks, largely due to its modeling distinctiveness that contributes to constructing and reweighing a relation among multi-perspective information. Inspired by this, we hope to introduce the knowledge enhancement into the task of video scene recognition to achieve an improvement. However, due to the conflict between the domain specificity and generality of knowledge enhancement, and the lack of general and excellent ways of fusing multi-perspective information with knowledge in the field of video understanding, how to leverage the external knowledge in our model is the second challenge. Moreover, knowledge-enhanced models are usually large-scale with low computational efficiency, and we expect to balance the efficiency and the additional time-consuming of introducing knowledge. Hence, it brings the third challenge that how to make the model as efficient as possible.

In recent years, there has been an increasing amount of literature on video understanding [6, 15, 27]. These works are usually based on spatio-temporal relationship modeling, and build models for video from multiple perspectives.

However, due to paying too much attention to the temporal information of videos, these works may ignore or lose the non-temporal information to varying degrees. In addition, several studies have revealed that external knowledge is beneficial for obtaining better performance [9, 10, 25]. Specifically, these studies focus on linking linguistic tokens to entities in knowledge graphs, and enhancing reasoning based on the neighborhood of entities, while lacking attention to vision due to scarce general visual entity linking. Meanwhile, numerous studies have also focused on the reasoning efficiency [5, 11, 15]. These studies employ various methods to distill models, which give us great inspiration.

In this paper, we propose a novel two-stream framework to learn video representations from multiple perspectives. Additionally, we design a knowledge-enhanced feature fusion method to integrate the multi-perspective information, which makes it natural to introduce knowledge into label recognition tasks. Finally, the application of knowledge distillation is employed to fuse the multiple representations, hoping to achieve high performance with low computational costs.

Our main contributions are as follows:

- We propose a novel two-stream framework to model videos from temporal and non-temporal perspectives, and integrate the two perspectives through knowledge distillation in order to better comprehend videos while maintaining efficiency.
- We design a knowledge-enhanced feature fusion method to leverage external knowledge to better integrate features non-temporally, and introduce knowledge into the scene label prediction task naturally while additionally gaining label scalability as a byproduct.
- We quantitatively evaluated our model on a real-world dataset. Experimental results demonstrate the effectiveness of our model.

The rest of this paper is structured as follows. In Section 2, we briefly review the related literature. In Section 3, we detail our proposed model, followed by experimental results and analyses in Section 4. We finally conclude the work in Section 5.

2. RELATED WORK

In this section, we mainly review the studies that are most related to our work, including video representation, scene recognition, and knowledge-enhanced learning.

2.1. Video Representation

Obtaining video representation is essential and indispensable for video analytics, while for obtaining video representation, spatio-temporal modeling of videos is an important step [16, 17, 26]. There are many attempts have been made to model videos from spatial and temporal perspectives [6, 15, 19, 29]. Hu et al. [6] proposed a hierarchical temporal method to construct the temporal structure at frame-level and object-level successively, and extract pivotal information effectively from global to local, which improves the model capacity of recognizing fine-grained objects and actions. Pan et al. [15] proposed a novel spatiotemporal graph network to explicitly exploit the spatiotemporal object interaction, which is crucial for video understanding and description. Besides, Shi et al. [19] proposed to learn the semantic concepts explicitly and design a temporal alignment mechanism to better align the video and transcript. These studies give us a lot of inspiration, however, they are too focused on the impact of temporal information on video comprehension due to task constraints, while ignoring non-temporal information to some extent.

2.2. Scene Recognition

Scene recognition is a task to develop robust and reliable models for the automatic recognition of what scenes are described by visual information [13]. Early research on scene recognition mainly focus on separate images [30], while scholars' attention naturally turn towards scene recognition from videos [7, 28]. Zhou et al. [30] described the Places Database, and provided scene recognition convolutional neural networks as baselines. Jiang et al. [7] proposed a novel framework, which jointly utilized multi-platform data, object-scene deep features and the hierarchical venue structure prior for scene category prediction from videos. Zhang et al. [28] proposed a Hybrid-Attention Enhanced Two-Stream Fusion Network for the task of video scene label prediction, and develops a novel Global-Local Attention Module, which can be inserted into neural networks to generate enhanced visual features from video content. Most recent studies identify scenes only from visual or textual information in a temporal perspective, ignoring the valuable information hidden in single frames, while several earlier studies only recognize scenes for separate images in nontemporal perspective. In this paper, we argue these two perspectives are both meaningful for this task and complementary to each other.

2.3. Knowledge-enhanced Learning

Knowledge-enhanced learning is increasingly attracting more attention from researchers in recent years. To make models better mine the hidden information in data, many scholars tried to introduce different types of additional information into their methods [5,9,11,15,25,27,31]. This additional information is considered as knowledge, and is usually related to knowledge graphs [9,25], transferred knowledge [5,11,15], and specifically defined knowledge [27,31].

When it comes to knowledge, scholars often refer to knowledge graphs and the embedding representations of the nodes and edges in knowledge graphs. Several studies [9,10,20,24,25] have been carried out on knowledge graphs and verified that the pretrained embeddings from knowledge graphs are helpful to reasoning. Liu et al. [9] proposed a knowledge-enabled language representation model with knowledge graphs, in which triples are injected into the sentences as domain knowledge. Xiong et al. [25] introduced a novel method to represent queries and documents in the entity space, and rank documents based on their semantic relatedness to queries. These studies fully explore the node embeddings and edges in knowledge graphs, while they mainly focus on text modality. And in our work, we leverage the pretrained knowledge embeddings to guide cross-modality fusion instead of single-modal reasoning.

Compared with knowledge graphs focusing on mining the valuable information hidden in nodes and edges in themselves, transferred knowledge focus on the transmission and sharing of known valuable information (e.g., distributions). The transferred knowledge is usually related to the pretrained models [3,4], especially pretrained language models [3], and the transferred distributions are common in the task of knowledge distillation [5, 11, 15, 29]. Pan et al. [15] proposed a novel spatio-temporal graph network and designed a two-branch framework with an object-aware knowledge distillation mechanism. Zhang et al. [29] proposed a novel teacher-recommended learning method that introduces external language model to guide the main model to learn abundant linguistic knowledge. In these methods, for better reasoning, knowledge is transferred between multiple modules that perform the same task in different ways, and this idea is adopted in our work.

The specifically defined knowledge in certain scenarios also appears in numerous recent studies [27, 31]. Zhang et al. [27] proposed to narrate the user-preferred product characteristics depicted in user-generated product videos, and proposed a novel framework to perform knowledgeenhanced spatio-temporal inference on product-oriented video graphs. Zhu *et al.* [31] introduced knowledge modality in multi-modal pretraining to correct the noise and supplement the missing image and text modalities. The aforementioned methods introduce several types of additional associated information, and leverage the specifically defined knowledge to guide the learning of models. However, the introduced knowledge also brings barriers to these methods, because the knowledge and methods usually target specific datasets, and are not easy to be extended to other tasks. In our work, though we similarly introduce additional specifically defined information (i.e., keywords) as knowledge, yet the introduced keywords are easily obtained from text descriptions, making our method extendable and adjustable.

3. OUR PROPOSED METHOD

3.1. Overview

3.1.1 Problem setting.

Before describing our method, we briefly introduce the problem setting first. Formally, let \mathcal{V} , \mathcal{L} , and \mathcal{G} denote the set of videos, the set of scene labels, and the external knowledge graph respectively. Each video $v \in \mathcal{V}$ contains textual descriptions t and a sequence of RGB frames. Scene labels \mathcal{L} belong to a two-level scene hierarchy, and each video is associated with a set of paths $\mathcal{P}_v = \{p_1, p_2, ..., p_{|\mathcal{P}_v|}\}$ on this hierarchy, where $p_l = \{p_l^1, p_l^2 \mid p_l^1, p_l^2 \in \mathcal{L}\}$ for $p_l \in \mathcal{P}_v$, and p_l^1 is the parent scene label of p_l^2 . Our goal is to learn a video scene recognition model, which could recognize suitable scene labels p based on the video frame sequences and the associated text descriptions.

3.1.2 Model overview.

Temporal information can well describe what happens in videos, while non-temporal information can well reflect several moments in videos. We argue that these two perspectives are both meaningful for this task and complementary to each other. Meanwhile, external introduced knowledge can also promote the comprehension of videos. In order to make full use of these kinds of information, we propose a novel two-stream framework to model video representations from the two perspectives, i.e. temporal and non-temporal perspectives, and integrate these two perspectives in an end-to-end manner by self-distillation. Besides, in the non-temporal module, we design a knowledge-enhanced feature fusion and label prediction method that contributes to naturally introducing knowledge into the task of video scene recognition.

3.2. Temporal Feature Learning

The temporal feature learning module is used to model videos from a temporal perspective, including temporal modeling and hierarchical multi-label prediction.

3.2.1 Temporal modeling.

In this module, we employ Multimodal Bitransformers [8] as the backbone, and utilize Transformer [23] that is effective in various tasks in recent years as the feature encoder. Specifically, for each video v, we uniformly sample $N_{\rm f}$ frames as keyframes, and utilized the pretrained ResNet [4] to extract the frame-level 2D feature $f_j^{\rm 2D}$ of each keyframe, where $j \in [1, N_{\rm f}]$. We then collect $N_{\rm c}$ consecutive frames with each sampled keyframe as center, and utilized the pretrained I3D [2] to extract the frame-level 3D features $f_j^{\rm 3D}$. For the associated textual descriptions, we utilize the pre-

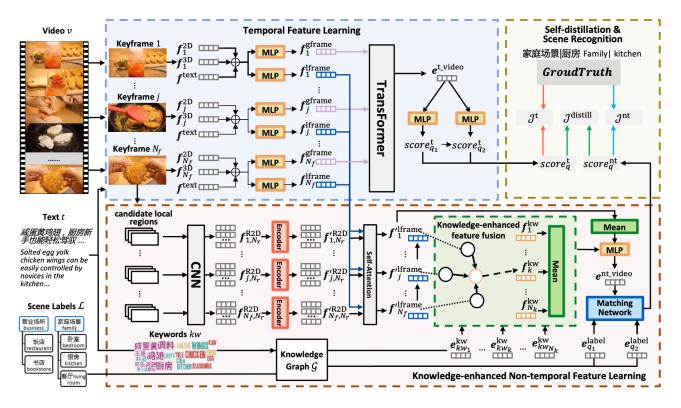


Figure 2. An overview of our proposed framework for scene recognition using knowledge-enhanced multi-perspective video representation learning.

trained BERT [3] to extract the video-level textual features $f^{\rm text}$.

After extracting the aforementioned features, we concatenate them to obtain the frame-level global features $f_j^{\rm gframe}$ and local features $f_j^{\rm lframe}$ as follows:

$$e_j^{\text{con}_{,\tau}} = (f_j^{\text{2D}} \oplus f_j^{\text{3D}} \oplus f^{\text{text}}),$$
 (1)

$$e_j^{\text{ref}_{_{\scriptscriptstyle \mathcal{I}}}} = \boldsymbol{W}_2^{\tau} \, \delta(\boldsymbol{W}_1^{\tau} e_j^{\text{con}_{_{\scriptscriptstyle \mathcal{I}}}} + \boldsymbol{b}_1^{\tau}) + \boldsymbol{b}_2^{\tau},$$
 (2)

$$\mathbf{f}_{j}^{\tau} = Norm(Dropout(\mathbf{e}_{j}^{\text{ref}.\tau}) + \mathbf{W}_{3}^{\tau} \mathbf{e}_{j}^{\text{con}.\tau}), \quad (3)$$
$$\tau \in \{\text{gframe}, \text{lframe}\},$$

where \oplus indicates concatenation, $\boldsymbol{e}_{j}^{\mathrm{con}}$ and $\boldsymbol{e}_{j}^{\mathrm{ref}}$ indicate the concatenated and the refined features respectively, $\delta(\cdot)$ indicates the activation function, $\boldsymbol{W}_{1}, \boldsymbol{W}_{2}, \boldsymbol{b}_{1}$ and \boldsymbol{b}_{2} denote the weight matrices and bias vectors in fully connected layers, \boldsymbol{W}_{3} denotes the residual transformation matrices, and Norm refers to the operation of layer normalization.

In order to better model the temporal aspects of video, we send the special tag [CLS] and the frame-level global features $f_j^{\rm gframe}$ into the self-attention layers of the Transformer encoder with position embeddings. After the encoding, the Transformer encoder generates a sequence of outputs, and we denote the generated output corresponding to the special tag [CLS] as the video-level temporal features

 $e^{\mathrm{t_video}}$ as follows:

$$O = [o_0, o_1, ..., o_{N_f}]$$

= Transformer($[[CLS] + pos_0,$ (4)

$$egin{aligned} oldsymbol{f}_{1}^{ ext{gframe}} + oldsymbol{pos}_{1}, ..., oldsymbol{f}_{N_f}^{ ext{gframe}} + oldsymbol{pos}_{N_f}]), \ e^{ ext{t-video}} = oldsymbol{o}_{0}, \end{aligned}$$

where O indicates the output sequence, pos_j ($j \in [0, N_{\rm f}]$) indicates the position embedding, and $N_{\rm f}$ denotes the number of sampled keyframes.

3.2.2 Hierarchical multi-label prediction.

After obtaining the video-level temporal features $e^{t\text{-}v\text{ideo}}$, we employ multi-layer perceptrons to obtain the basic scores of labels $score_i^{t}$ for the i^{th} -level scene hierarchical layer. Besides, for each label q_1 in the 1^{st} -level scene hierarchical layer, we make the refined scores $score_{q_1}^t$ of it the same as its basic score. And for each label q_2 in the 2^{rd} -level layer, we obtain the refined scores $score_{q_2}^t$ of it by summing the basic scores of itself and its parent label μ_{q_2}

as follows:

$$score'_i^{\rm t} = \boldsymbol{W}_5^i \, \delta(\boldsymbol{W}_4^i \boldsymbol{e}^{\rm t.video} + \boldsymbol{b}_4^i) + \boldsymbol{b}_5^i,$$
 (6)

$$score_{q_1}^{t} = score_{1,q_1}^{\prime t}, \qquad (7)$$

$$score_{q_1}^t = score_{1,q_1}^{t}, \qquad (7)$$

 $score_{q_2}^t = score_{1,\mu_{q_2}}^{t} + score_{2,q_2}^{t}, \qquad (8)$

where $score_i^{\mathrm{t}} \in \mathbb{R}^{|\mathcal{L}_i|}$ and $|\mathcal{L}_i|$ denote the scores and the number of the labels in the $i^{
m th}$ -level scene hierarchical layer respectively, μ_{θ} denotes the parent label of θ in scene hierarchy, $\delta(\cdot)$ indicates the activation function, $\boldsymbol{W}_4^i \in \mathbb{R}^{d_{\mathrm{emb}} \times d_{\mathrm{emb}}}$ and $\boldsymbol{W}_5^i \in \mathbb{R}^{|\mathcal{L}_i| \times d_{\mathrm{emb}}}$ denote weight matrices in fully connected layers, $\pmb{b}_4^i \in \mathbb{R}^{d_{\mathrm{emb}}}$ and $\pmb{b}_5^{ar{i}} \in \mathbb{R}^{|\mathcal{L}_i|}$ denote bias vectors, and d_{emb} denotes the dimensions of the video-level temporal features.

And then we employ the Multi-label Cross Entropy Loss [21, 22] to calculate the loss for the i^{th} -level scene hierarchical layer as follows:

$$\mathcal{J}_{i}^{t} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \{log[1 + \sum_{q_{i}^{-} \notin \mathcal{P}_{v}^{i}} exp(score_{q_{i}^{-}}^{t})] + log[1 + \sum_{q_{i}^{+} \in \mathcal{P}_{v}^{i}} exp(-score_{q_{i}^{+}}^{t})]\},$$
(9)

where $|\mathcal{V}|$ denotes the number of videos, q_i^+ and q_i^- denotes the positive and the negative scene labels in the i^{th} -level scene hierarchy respectively, and \mathcal{P}_{v}^{i} denotes the annotated scene labels of video v in the i^{th} -level scene hierarchy.

Finally, the final temporal objective function can be formulated as follows:

$$\mathcal{J}^{t} = \beta_1^{\text{level}} \mathcal{J}_1^{t} + \beta_2^{\text{level}} \mathcal{J}_2^{t}, \tag{10}$$

where β_1^{level} and $\beta_2^{rmlevel}$ are hyper-parameters as coefficients of \mathcal{J}_1^t and \mathcal{J}_2^t .

3.3. Knowledge-enhanced Non-temporal Feature Learning

The knowledge-enhanced non-temporal feature learning module is used to model videos from a knowledgeenhanced non-temporal perspective. We first obtain framelevel local features by fusing candidate local regions, and then introduce external knowledge to perform and enhance the video feature fusion and scene prediction.

3.3.1 Frame-level local feature fusion.

In this module, we employ the pretrained Faster-RCNN [18] to detect the candidate local regions, extract their 2D features with the pretrained ResNet, and denote the 2D feature of the m-th candidate regions of the j-th keyframe as $m{f}_{j,m}^{\mathrm{R2D}}$

Then we send all the candidate region features detected in each keyframe directly into the self-attention layers of

the Transformer encoder. We employ this way to complete inner-frame reasoning among the candidate regions detected in the same keyframe, and obtain the enhanced feature $f'_{j,m}^{\text{R2D}} \in f'_{j}^{\text{R2D}}$ of each candidate region as follows:

$$\boldsymbol{f'}_{j}^{\mathrm{R2D}} = \mathrm{Transformer}([\boldsymbol{f}_{j,1}^{\mathrm{R2D}}, \boldsymbol{f}_{j,2}^{\mathrm{R2D}}, ..., \boldsymbol{f}_{j,N_{\mathrm{r}}}^{\mathrm{R2D}}]), \quad (11)$$

where $N_{\rm r}$ indicates the number of candidate regions extracted in the same keyframe.

After that, we leverage the frame-level local features f_i^{lframe} obtained in the previous section as queries, perform self-attention operations on the enhanced candidate region feature $e_{j,m}^{\rm R2D} \in e_j^{\rm R2D}$, and denote the obtained fused frame-level local features $f_j'^{\rm lframe}$ as follows:

$$Query_j = f_j^{\text{lframe}} W_Q^{\text{local}},$$
 (12)

$$Key_{j,m} = f'_{j,m}^{\text{R2D}} W_{\text{K}}^{\text{local}},$$
 (13)

$$Key_{j,m} = f'_{j,m}^{\text{R2D}} W_{\text{K}}^{\text{local}},$$
 (13)
 $Value_{j,m} = f'_{j,m}^{\text{R2D}} W_{\text{V}}^{\text{local}},$ (14)

$$\alpha_{j,m}^{\text{frame}} = \operatorname{softmax}(\frac{Query_{j,m}Key_{j,m}}{\sqrt{d_{\text{Key}}}}), \quad (15)$$

$$f'_{j}^{\text{lframe}} = \sum_{m \in [1,M]} \alpha_{j,m}^{\text{frame}} Value_{j,m},$$
 (16)

where $m{W}_{
m Q}^{
m local}, m{W}_{
m K}^{
m local}, m{W}_{
m V}^{
m local}$ indicate the weight matrices corresponding to the queries, keys and values of the selfattention module, and $d_{\rm Key}$ represents the dimensions of the key vectors.

3.3.2 Knowledge-enhanced video feature fusion.

In this module, we leverage the entity embeddings pretrained from knowledge graphs \mathcal{G} as external knowledge, and denote G_{θ} as the pretrained embeddings corresponding to the token θ . In addition, we also employ the embeddingbased [3, 14] unsupervised keyword extraction algorithm to extract N_k keywords $kw_k \in \mathcal{K}_v \subseteq \mathcal{G}$ from the associated text descriptions of video v.

In terms of feature fusion, inspired by NetVLAD [1], we design a knowledge-enhanced feature fusion method. Unlike NetVLAD, which directly learn a set of parameters for each cluster center and clusters features based on the measured distances, we employ shared parameter modules to generate the required parameters based on the features of the clustering target (i.e., the pretrained word embeddings of keywords) as follows:

$$\varphi_{\mathbf{w}}(\theta) = \boldsymbol{W}_{2}^{\mathbf{w}} \, \delta(\boldsymbol{W}_{1}^{\mathbf{w}} \theta + \boldsymbol{b}_{1}^{\mathbf{w}}) + \boldsymbol{b}_{2}^{\mathbf{w}}, \tag{17}$$

$$\varphi_c(\theta) = \boldsymbol{W}_2^c \, \delta(\boldsymbol{W}_1^c \theta + \boldsymbol{b}_1^c) + \boldsymbol{b}_2^c, \tag{18}$$

$$\varphi_{\mathbf{z}}(\theta) = \mathbf{W}_{2}^{\mathbf{z}} \, \delta(\mathbf{W}_{1}^{\mathbf{z}} \theta + \mathbf{b}_{1}^{\mathbf{z}}) + \mathbf{b}_{2}^{\mathbf{z}}, \tag{19}$$

where $\boldsymbol{W}_{1}^{\tau}$, $\boldsymbol{W}_{2}^{\tau}$, $\boldsymbol{b}_{1}^{\tau}$ and $\boldsymbol{b}_{2}^{\tau}$ ($\tau \in \{\mathrm{w,c,z}\}$) denote weight matrices and bias vectors in fully connected layers, $\varphi_{\rm w}(\theta)$

and $\varphi_c(\theta)$ denote the functions which can generate parameters to measure distances, $\varphi_{z}(\theta)$ denotes the function which is usd to generate the representations of the cluster centers, and $\delta(\cdot)$ indicates the activation function.

Then we measure the similarity between different framelevel local features and keyword semantics based on the learned parameters, and weighted summing the difference between the frame-level local features and the obtained representations of the keyword cluster centers. Through this design, we can fuse the features of the sampled frames to varying degrees based on the semantics of specific keywords, and obtain the knowledge-enhanced keyword-level non-temporal feature f_k^{kw} as follows:

$$\alpha_{j,k}^{\text{kw}} = \operatorname{softmax}(\varphi_{\text{w}}(G_{kw_k})f_j^{\prime \text{lirame}} + \varphi_{\text{c}}(G_{kw_k})), \quad (20)$$

$$\alpha_{j,k}^{\text{kw}} = \operatorname{softmax}(\varphi_{\text{w}}(\boldsymbol{G}_{kw_k})\boldsymbol{f}'_{j}^{\text{lframe}} + \varphi_{\text{c}}(\boldsymbol{G}_{kw_k})), \quad (20)$$
$$\boldsymbol{f}_{k}^{\text{kw}} = \sum_{j \in [1, N_f]} \alpha_{j,k}^{\text{video}}(\boldsymbol{f}'_{j}^{\text{lframe}} - \varphi_{\text{z}}(\boldsymbol{G}_{kw_k})), \quad (21)$$

where kw_k represents the k-th extracted keyword of video $v, N_{\rm f}$ denotes the number of the sampled keyframes of each video, and G_{θ} denotes the operation of obtaining the pretrained word embedding corresponding to the specific linguistic token θ .

After that, we utilize mean pooling to fuse multiple knowledge-enhanced keyword-level non-temporal features of the same video, and leverage the mean feature of the fused frame-level local features, to receive the knowledgeenhanced video-level non-temporal feature $e^{\mathrm{nt_video}}$ for each video as follows:

$$\boldsymbol{e}^{\prime \text{nt_video}} = \frac{1}{|\mathcal{K}|} \sum_{k \in [1, |\mathcal{K}|]} \boldsymbol{f}_k^{\text{kw}} + \frac{1}{N_f} \sum_{j \in [1, N_f]} \boldsymbol{f}_j^{\prime \text{lframe}}, \quad (22)$$

$$e^{\text{nt_video}} = W^{\text{nt}} e'^{\text{nt_video}} + b^{\text{nt}},$$
 (23)

where $oldsymbol{W}^{\mathrm{nt}}$ and $oldsymbol{b}^{\mathrm{nt}}$ denote weight matrices and bias vectors in the fully connected layer, $|\mathcal{K}|$ denotes the number of the extracted keywords of videos, and N_f denotes the number of the sampled keyframes for each video.

3.3.3 Knowledge-enhanced multi-label scene prediction

In order to make better use of knowledge for scene prediction, we design shared matching networks to calculate the basic matching scores $score'^{\rm nt}_{v,q}$ between videos v and the i^{th} -level scene label q_i , based on the knowledge-enhanced video-level non-temporal featrue $oldsymbol{e}_v^{ ext{nt_video}}$ and the scene label representations $e_{q_i}^{\mathrm{label}}$. After that, we obtain the refined scores $score_{q_i}^{\mathrm{nt_video}}$ using approaches similar to those in the hierarchical multi-label prediction in the temporal module

as follows:

$$e'_{v,i}^{\text{nt_video}} = \boldsymbol{W}_7^i \delta(\boldsymbol{W}_6^i \boldsymbol{e}_v^{\text{nt_video}} + \boldsymbol{b}_6^i) + \boldsymbol{b}_7^i,$$
 (24)

$$e'_{q_i}^{\text{label}} = \boldsymbol{W}_9^i \, \delta(\boldsymbol{W}_8^i e_{q_i}^{\text{label}} + \boldsymbol{b}_8^i) + \boldsymbol{b}_9^i, \qquad (25)$$

$$score'_{v,q_i}^{\text{nt}} = e'_{v,i}^{\text{nt-video}} \circ e'_{q_i}^{\text{label}}, \qquad (26)$$

$$score'_{v,q_i}^{\text{nt}} = e'_{v,i}^{\text{nt_video}} \circ e'_{q_i}^{\text{label}},$$
 (26)

$$score_{v,q_1}^{\text{nt}} = score_{v,q_1}^{\prime \text{nt}},$$
 (27)

$$score_{v,q_2}^{\text{nt}} = score_{v,\mu_{q_2}}^{'\text{nt}} + score_{v,q_2}^{'\text{nt}},$$
 (28)

where $\boldsymbol{W}_{\tau}^{i}$ and $\boldsymbol{b}_{\tau}^{i}$ $(i \in \{1,2\}, \tau \in \{6,7,8,9\})$ denote weight matrices and bias vectors in fully connected layers, $\delta(\cdot)$ indicates the activation function, \circ represents the inner product operation, and μ_{θ} denotes the parent label of θ in scene hierarchy.

For those scene labels whose corresponding entity can be found in the knowledge graph \mathcal{G} , we directly utilize their corresponding pretrained entity embeddings G_q as the scene representations e_q^{label} . And for the remaining unmatched scene labels, we initialize their embeddings randomly and update them when training the network. In this way, we also receive an additional benefit, which is the scalability of labels. Specifically, the scalability of labels means that, when we need to predict an unseen label that is not in the original label list but is an entity in knowledge graphs, we do not need to retrain the model, and can perform inferences directly.

Similar to the temporal module, we also employ the Multi-label Cross Entropy Loss to calculate the loss $\mathcal{J}_i^{\text{nt}}$ for each scene hierarchical layer, and the final non-temporal objective function can be formulated as follows:

$$\mathcal{J}^{\text{nt}} = \beta_1^{\text{level}} \mathcal{J}_1^{\text{nt}} + \beta_2^{\text{level}} \mathcal{J}_2^{\text{nt}}, \tag{29}$$

where β_1^{level} and β_2^{level} are hyper-parameters as coefficients of $\mathcal{J}_1^{\mathrm{nt}}$ and $\mathcal{J}_2^{\mathrm{nt}}$.

3.4. Self-distillation and Scene Recognition

In the aforementioned sections, we model videos from two perspectives (i.e., the temporal perspective and the knowledge-enhanced non-temporal perspective), and obtain scene label scores from both perspectives.

To enable the two modules (i.e., temporal module and the knowledge-enhanced non-temporal module) to learn information separately, and integrate the learned information with each other, we employ Euclideandistance to measure the difference between the two groups (i.e., the temporal perspective and the knowledge-enhanced non-temporal perspective) of scene label scores in each scene hierarchical layer. We obtain the distillation loss $\mathcal{J}^{\mathrm{distill}}$ according to the obtained Euclidean distances $\mathcal{J}_i^{ ext{distill}}$ for the i^{th} -level scene hierarchy as follows:

$$\mathcal{J}_{i}^{\text{distill}} = \frac{1}{|\mathcal{V}|} sqrt \left[\sum_{q_{i} \in \mathcal{L}_{i}} (score_{q_{i}}^{t} - score_{q_{i}}^{\text{nt}})^{2} \right], \quad (30)$$

$$\mathcal{J}^{\text{distill}} = \beta_1^{\text{level}} \mathcal{J}_1^{\text{distill}} + \beta_2^{\text{level}} \mathcal{J}_2^{\text{distill}}, \qquad (31)$$

where β_1^{level} and β_2^{level} are hyper-parameters as coefficients of $\mathcal{J}_1^{\mathrm{distill}}$ and $\mathcal{J}_2^{\mathrm{distill}}$, and \mathcal{L}_i denotes the set of the scene labels in the i^{th} -level hierarchical layer.

In the end, the final objective function is defined as:

$$\mathcal{J} = \beta^{t} \mathcal{J}^{t} + \beta^{nt} \mathcal{J}^{nt} + \beta^{distill} \mathcal{J}^{distill}, \qquad (32)$$

where β^t , β^{nt} and $\beta^{distill}$ are hyper-parameters as coefficients of \mathcal{J}^t , \mathcal{J}^{nt} and $\mathcal{J}^{distill}$.

Moreover, to balance the performance and efficiency of the model, inspired by previous work [5, 11, 15], the two modules are both utilized to participate in the training, but only the temporal module is employed for reasoning. When reasoning, given videos and the associated textual descriptions as queries, the model will first calculate the scene label scores through the temporal module, then take out the scene labels with the Top-K scores or the scores that exceed a specific threshold, and return them to users.

4. EXPERIMENTS

In this section, we conduct experiments on a real-world dataset to evaluate the performance of our proposed method.

4.1. Dataset

We evaluate our model on one of the largest available video scene datasets, which is called the Koubei dataset. The Koubei Dataset contains metadata from Koubei Platform¹, including 63,977 videos with associated textual descriptions, and manually annotated hierarchical scene labels, where each video can correspond to multiple scene labels. Besides, the scene label hierarchy has 6 1^{st} -level labels and 320 2^{rd} -level labels.

We randomly split the Koubei dataset into training, validation, and testing sets. Specifically, we randomly sample 1,000 videos into the validation and 1,000 videos into the testing sets, respectively, and the remaining 61,977 videos are utilized as the training set.

4.2. Evaluation Protocol and Parameters Settings

We evaluate the performance of different models using F1 score and RP@90% as the evaluation metrics. RP@90% denotes the proportion of labelled videos when we add threshold constraints to make Accuracy reaches 90%, and this metric is widely employed in commercial products.

In our experiments, we set the number of the sampled keyframes N_f as 12, the number of consecutive frames for 3D features N_c as 16, and the maximum number of extracted keywords N_r as 10. For simplicity, We set all the hyper-parameters β^{t} , β^{nt} , $\beta^{distill}$, $\beta^{distill}_{1}$, $\beta^{distill}_{2}$, β^{level}_{1} , β_2^{level} to the same value 1. We introduce *ConceptNet* as the external knowledge graph, and the pretrained entity embeddings are provided by ConceptNet Numberbatch [20]. We utilize GeLU as the activation function, and employ dropout with 50% keep probability, weight decay, and early stopping to alleviate overfitting. To train our proposed model, we randomly initialize model parameters with a Gaussian distribution, and utilize AdamW [12] algorithm for optimization. We further restrict the dimensions of the final representation vector of each video to be the same for fair comparisons. We have tried different parameter settings, including the batch size of {8, 16, 32}, the latent feature dimension of {192, 384, 768}, the learning rate of {1e-1, 3e-4, 1e-4, 3e-5, 1e-5. As the findings are consistent across the dimensions of latent vectors, if not specified, we only report the results based on the dimension of 768, which gives relatively good performance.

4.3. Baselines

To evaluate the effectiveness of our model, we compare our proposed method with several state-of-the-art baselines. Specifically, the original task of the first four models is not video scene recognition, we adjust these models and perform the task of scene prediction based on the obtained video representations using them. The latter two models which are originally designed for scene recognition are directly utilized in experiments.

- MMBT [8]: introduces a supervised multi-modal bi-Transformer that jointly finetunes uni-modally pretrained text and image encoders. Besides, this model is the backbone of the temporal module of our model.
- LSCTA [19]: employs a Transformer model as the backbone, and develops a framework to align sequences in different modalities to capture information.
- HGLTM [6]: proposes a hierarchical model which can construct the temporal structure at frame-level and object-level successively, and extract pivotal information effectively from global to local, which improves the model capacity.
- STGK [15]: proposes a novel spatio-temporal graph network to explicitly exploit the spatio-temporal object interaction, which is crucial for video understanding.
- HCMFL [7]: develops a Hierarchy-dependent Crossplatform Multi-view Feature Learning framework, which jointly utilizes multi-platform data, object-scene

¹www.koubei.com

Table 1. Performance comparison among our model and all fully-trained baselines using the metrics RP@90% and F1 score.

Model	RP@90%	F1 score
MMBT	0.521	0.720
LSCTA	0.324	0.706
HGLTM	0.434	0.719
STGK	0.452	0.707
HCMFL	0.445	0.735
HETFN	0.483	0.710
Ours-S2	0.536	0.726
Ours-S2 ^{w/o-hier}	0.511	0.731
Ours-S2 ^{w/o-know&w-temp}	0.511	0.709
Ours-S2 ^{w/o-know}	0.504	0.707
Ours-S2 ^{w/o-kw}	0.531	0.706
OurModel	0.561	0.735

deep features, and the hierarchical structure prior for category prediction from videos.

- HETFN [28]: proposes a Hybrid-Attention Enhanced Two-Stream Fusion Network for the video recognition task, and develops a novel Global-Local Attention Module, which can be inserted into neural networks to generate enhanced visual features from video content.
- Ours-S2: This variant removes the Temporal Feature Learning module (S1) from the origin proposed model, and keeps the Knowledge-enhanced Non-temporal Feature Learning module (S2) to learn the representation of videos.
- Ours-S2^{w/o-hier}: This variant additionally removes the loss functions for the 1st-level scene hierarchical layer on the basis of Ours-S2.
- Ours-S2^{w/o-know&w-temp}: This variant replaces the frame-level local feature fusion and the knowledge-enhanced video feature fusion in the non-temporal module by *Transformer* on the basis of Ours-S2.
- Ours-S2^{w/o-know}: This variant drops the proposed knowledge-enhanced video feature fusion module, and utilizes the mean feature of the fused frame-level local features as the video-level non-temporal feature on the basis of Ours-S2.
- Ours-S2^{w/o-kw}: This variant replaces the employed pretrained keyword entity embeddings by an equal number of randomly initialized embeddings shared by all videos on the basis of Ours-S2.

4.4. Performances, Quantitative Analysis and Ablation Study

We evaluate all fully-trained models using the metrics F1 score and RP@90%, and report the results in Table 1. We

have the following observations with respect to our experimental results.

First, our proposed method achieves the best performance on Koubei Dataset using the metric RP@90% and F1 score, demonstrating the effectiveness of our model.

Second, compared with MMBT which is the backbone of the temporal module of OurModel, and models videos only from the temporal perspective, the whole OurModel achieves a 7.7% performance gain on RP@90% and 2.1% performance gain on F1 score, demonstrating the effectiveness of our proposed knowledge-enhanced non-temporal module. Besides, OurModel achieves better performance than separate stream modules MMBT and Ours-S2, verifying that the two perspectives (*i.e.*, the temporal and the knowledge-enhanced non-temporal perspectives) are both valuable for the task of scene recognition and complementary to each other.

Third, compared with Ours-S2^{w/o-hier}, Ours-S2 receives better performance, verifying that calculating loss function simultaneously for multi-level scene hierarchy can improve the performance of our model.

Moreover, Ours-S2 obtain better performance than Ours-S2 $^{\text{w/o-know}\&\text{w-temp}}$ and Ours-S2 $^{\text{w/o-know}}$, demonstrating the effectiveness of the proposed frame-level local feature fusion and knowledge-enhanced video feature fusion.

Finally, compared with Ours-S2^{w/o-kw}, Ours-S2 achieves better performance, verifying that the introduced external knowledge is valuable to our proposed model.

5. Conclusions and Future Work

With the explosive growth of video data in real-world applications, a comprehensive representation of videos becomes increasingly important. In this paper, we address the problem of video scene recognition, whose goal is to learn a high-level video representation to classify scenes in videos. In this paper, we propose a novel two-stream framework to model video representations from the temporal and knowledge-enhanced non-temporal perspectives, and integrate these two perspectives in an end-to-end manner by self-distillation. Besides, we design a knowledge-enhanced feature fusion and label prediction method that contributes to naturally introducing knowledge into the task of video scene recognition. We evaluated our model for scene recognition on a real-world dataset, and the experimental results demonstrate the effectiveness of our proposed model. In addition, we also conducted ablation studies, which demonstrated the effectiveness of our proposed temporal and nontemporal two-stream framework and knowledge-enhanced feature fusion method, respectively. In future work, we may pay more attention to the utilization of knowledge graphs, and try to leverage edge information in knowledge graphs to assist reasoning to enhance video understanding.

References

- [1] Relja Arandjelovic, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Netvlad: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5297–5307. IEEE Computer Society, 2016. 5
- [2] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, pages 4724–4733. IEEE Computer Society, 2017. 3
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics, 2019. 3, 4, 5
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 770–778. IEEE Computer Society, 2016. 3
- [5] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. CoRR, abs/1503.02531, 2015. 2, 3, 7
- [6] Yaosi Hu, Zhenzhong Chen, Zheng-Jun Zha, and Feng Wu. Hierarchical global-local temporal modeling for video captioning. In *Proceedings of the ACM International Conference on Multimedia*, MM, pages 774–783. ACM, 2019. 1, 2, 7
- [7] Shuqiang Jiang, Weiqing Min, and Shuhuan Mei. Hierarchydependent cross-platform multi-view feature learning for venue category prediction. *IEEE Trans. Multim.*, 21(6):1609–1619, 2019. 2, 7
- [8] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. In *Visually Grounded Interaction* and Language (ViGIL), NeurIPS Workshop, 2019. 3, 7
- [9] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-BERT: enabling language representation with knowledge graph. In The AAAI Conference on Artificial Intelligence, AAAI, The Innovative Applications of Artificial Intelligence Conference, IAAI, The AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI, pages 2901–2908. AAAI Press, 2020. 2,
- [10] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL*, pages 2395–2405. Association for Computational Linguistics, 2018. 2, 3
- [11] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Represen*tations, ICLR, Conference Track Proceedings, 2016. 2, 3,

- [12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations, ICLR*. OpenReview.net, 2019. 7
- [13] Alina Matei, Andreea Glavan, and Estefanía Talavera. Deep learning for scene recognition from visual data: A survey. In *Hybrid Artificial Intelligent Systems*, *HAIS*, volume 12344 of *Lecture Notes in Computer Science*, pages 763– 773. Springer, 2020. 2
- [14] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, ICLR Workshop Track Proceedings, 2013. 5
- [15] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *IEEE/CVF Conference on Computer Vision* and Pattern Recognition, CVPR, pages 10867–10876. Computer Vision Foundation / IEEE, 2020. 1, 2, 3, 7
- [16] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *IEEE International Conference on Computer Vision*, *ICCV*, pages 5534–5542. IEEE Computer Society, 2017. 2
- [17] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. Learning spatio-temporal representation with local and global diffusion. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 12056–12065. Computer Vision Foundation / IEEE, 2019. 2
- [18] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems, pages 91–99, 2015. 5
- [19] Botian Shi, Lei Ji, Zhendong Niu, Nan Duan, Ming Zhou, and Xilin Chen. Learning semantic concepts and temporal alignment for narrated video procedural captioning. In Proceedings of the ACM International Conference on Multimedia, MM, pages 4355–4363. ACM, 2020. 2, 7
- [20] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pages 4444–4451. AAAI Press, 2017. 3,
- [21] Jianlin Su. Extend 'Softmax+Cross Entropy' to Multi-label Classification Problem, 2020. 5
- [22] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 6397–6406. Computer Vision Foundation / IEEE, 2020. 5
- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems, pages 5998–6008, 2017. 3
- [24] Chenyan Xiong, Jamie Callan, and Tie-Yan Liu. Word-entity duet representations for document ranking. In *Proceedings*

- of the International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 763–772. ACM, 2017. 3
- [25] Chenyan Xiong, Russell Power, and Jamie Callan. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the International Conference* on World Wide Web, WWW, pages 1271–1279. ACM, 2017. 2, 3
- [26] Yuan Yuan, Haopeng Li, and Qi Wang. Spatiotemporal modeling for video summarization using convolutional recurrent neural network. *IEEE Access*, 7:64676–64685, 2019.
- [27] Shengyu Zhang, Ziqi Tan, Jin Yu, Zhou Zhao, Kun Kuang, Jie Liu, Jingren Zhou, Hongxia Yang, and Fei Wu. Poet: Product-oriented video captioner for e-commerce. In *Proceedings of the ACM International Conference on Multimedia, MM*, pages 1292–1301. ACM, 2020. 1, 2, 3
- [28] Yanchao Zhang, Weiqing Min, Liqiang Nie, and Shuqiang Jiang. Hybrid-attention enhanced two-stream fusion network for video venue prediction. *IEEE Trans. Multim.*, 23:2917– 2929, 2021. 2, 8
- [29] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 13275–13285. Computer Vision Foundation / IEEE, 2020. 2, 3
- [30] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2018. 2
- [31] Yushan Zhu, Huaixiao Zhao, Wen Zhang, Ganqiang Ye, Hui Chen, Ningyu Zhang, and Huajun Chen. Knowledge perceived multi-modal pretraining in e-commerce. In *Proceedings of the ACM International Conference on Multimedia, MM*, pages 2744–2752. ACM, 2021. 2, 3