Hybrid-Attention Enhanced Two-Stream Fusion Network for Video Venue Prediction

Yanchao Zhang, Weiqing Min, *Member, IEEE*, Liqiang Nie, *Senior Member, IEEE*, and Shuqiang Jiang, *Senior Member, IEEE*

***Abstract*—Video venue category prediction has been drawing more attention in the multimedia community for various applications such as personalized location recommendation and video verification. Most of existing works resort to the information from either multiple modalities or other platforms for strengthening video representations. However, noisy acoustic information, sparse textual descriptions and incompatible cross-platform data could limit the performance gain and reduce the universality of the model. Therefore, we focus on discriminative visual feature extraction from videos by introducing a hybrid-attention structure. Particularly, we propose a novel Global-Local Attention Module (GLAM), which can be inserted to neural networks to generate enhanced visual features from video content. In GLAM, the Global Attention (GA) is used to catch contextual scene-oriented information via assigning channels with various weights while the Local Attention (LA) is employed to learn salient object-oriented features via allocating different weights for spatial regions. Moreover, GLAM can be extended to ones with multiple GAs and LAs for further visual enhancement. These two types of features respectively captured by GAs and LAs are integrated via convolution layers, and then delivered into convolutional Long Short-Term Memory (convLSTM) to generate spatial-temporal representations, constituting the content stream. In addition, video motions are explored to learn long-term movement variations, which also contributes to video venue prediction. The content and motion stream constitute our proposed Hybrid-Attention Enhanced Two-Stream Fusion Network (HA-TSFN). HA-TSFN finally merges the features from two streams for comprehensive representations. Extensive experiments demonstrate that our method achieves the state-of-the-art performance in the large-scale dataset Vine. The visualization also shows that the proposed GLAM can capture complementary scene-oriented and object-oriented visual features from videos. Our code is available at: https://github.com/zhangyanchao1014/HA-TSFN.**

I. Introduction

Recent years have witnessed the fast development of geographic location prediction in the multimedia community and beyond [1]–[7] for various applications, e.g., geographic information retrieval and image/video verification. One important branch is venue category prediction from videos, which is to predict the venue category based on one video, such as Restaurant and Gym. It is especially valuable for social media oriented searching and recommendation. For example, many users often upload videos without venue tagging to the social media network, such as Facebook and Vine when they dine out. If we can infer food-related venue categories from those uploaded videos, such as fast-food restaurant and coffee shop, we can mine their dietary habits, and further recommend restaurants or healthy recipes for dietary management [8]. Therefore we focus on video venue category prediction in this work because of its great potentials.

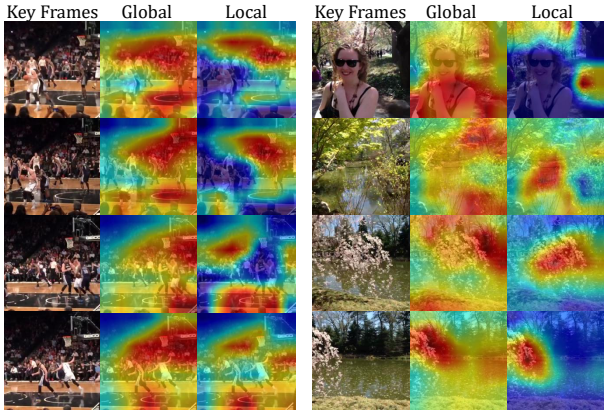


Fig. 1: The illustration of two videos from the Vine dataset. (a) The global attention learns the scene composed of the athletes, audience, basketball, basket, ground and their arrangements while the local attention mainly attends to key objects: athletes, audiences and ground. (b) The global attention studies the scene of the flowers, trees, lake, people and their position distributions while the local attention only focuses on the flowers and lake.

Although there are a lot of online video sharing platforms, few users upload videos with venue information. This incurs that the video dataset with venue annotations is in shortage. In order to tackle this problem, Zhang *et al.* [1] published the first large-scale video dataset Vine with venue category annotations. They also proposed a tree-guided multi-task multi-modal method to estimate venue categories. Later, based on this benchmark dataset, Nie *et al.* [2] enhanced acoustic modality via harnessing external sound knowledge to improve the prediction performance. Jiang *et al.* [3] exploited cross-platform data to improve the classification performance. Existing works mainly utilize the information from multiply modalities or

This work was supported by the Shandong Provincial Key Research and Development Program, 2019JZZY010118; the National Natural Science Foundation of China, 61972378, 61532018, U1936203, U19B2040; the Shandong Provincial Natural Science Foundation, ZR2019JQ23; the Innovation Teams in Colleges and Universities in Jinan, 2018GXRC014. Y. Zhang is with the School of Computer Science and Technology, Shandong University, Tsingtao, 266000, and also an intern with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China email: zyc1641299934@gmail.com. L. Nie is with the School of Computer Science and Technology, Shandong University, Tsingtao, 266000, China email: nieliqiang@gmail.com. W. Min and S. Jiang are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China, and also with University of Chinese Academy of Sciences, Beijing, 100049, China email: minweiqing@ict.ac.cn, sqjiang@ict.ac.cn.

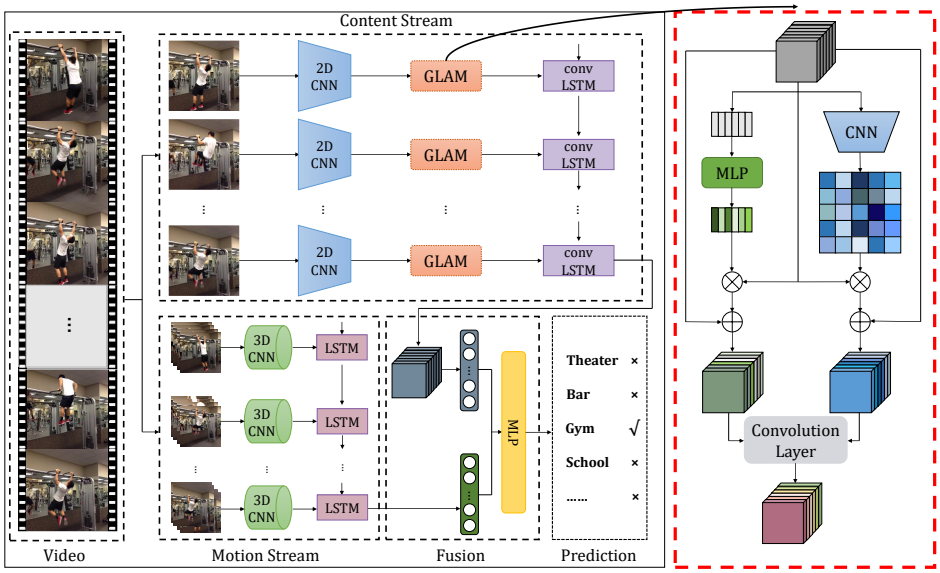


Fig. 2: The framework Hybrid-Attention Enhanced Two-Stream Fusion Network contains two streams: the content stream and motion stream. The content stream is augmented by GLAM. The convLSTM is utilized in the content stream and LSTM is utilized in the motion stream to learn long-term spatial-temporal information. The two streams are integrated to generate final comprehensive representations. The GLAM is illustrated in the red box.

other platform data to enhance video representations. However, noisy acoustic information, sparse textual descriptions and incompatible cross-platform data especially from social media networks could limit the performance gain and reduce the universality of the model.

To tackle above issues, we focus on exploring visual information from videos and are devoted to visual reinforcement. Considering that scene, object and motion information all contribute to inferring venue categories, we learn video representations by capturing scene-oriented and object-oriented features from content cues and catching motion-related features from motion cues. For example, in Fig. 1 (a), some athletes are playing basketball. This scene is composed of the athletes, audience, basketball, basket, ground and their arrangements while the key objects are the athletes, ground and audience, which jointly determine the category Basketball Stadium. The action playing basketball can supplement the former information. Similarly in Fig. 1 (b), the crucial objects flowers and lake together decide the video class Garden. Meanwhile the flowers, lake, trees, people and their location layouts constitute the garden scene. Besides, tree shaking and water waving are also helpful to infer the correct category.

Although some works [3], [9] have learned visual features from both perspectives of scene and object, they simply utilized a Place205-pretrained network to obtain scene-oriented features and an ImageNet-pretrained network to obtain object-oriented features. In contrast, we stress both scene-oriented and object-oriented features as well as explore their semantic representations explicitly by introducing a hybrid-attention mechanism. Particularly, we design a Global-Local Attention Module (GLAM) including a global attention and a local attention. The GLAM can be inserted into neural networks to enhance video content information. The global attention is to highlight contextual information (i.e., scene-related features) via assigning channels with various weights while the local attention is to emphasize salient areas (i.e., object-related features) via allocating different weights for spatial regions. Two visualization examples generated by the global and local attention are illustrated in Fig. 1. In addition, we utilize the 3D convolution network to capture motion information based on video clips.

We further propose a Hybrid-Attention Enhanced Two-Stream Fusion Network (HA-TSFN) to estimate video venue categories. As illustrated in Fig. 2, HA-TSFN mainly consists of two streams: (1) Content stream. 2D convolution networks (e.g., VGG [10] and ResNet [11]) are utilized to extract content information. In this stream, we propose a novel GLAM structure to enhance the visual representation ability of convolution neural networks (CNNs). The global attention catches contextual information while the local attention discovers salient regions. After crossing hybrid attentions, two kinds of feature maps are obtained, and then integrated via convolution layers. We further extend GLAM with multiple global and local attentions at any depth to generate more discriminative content features. (2) Motion stream. 3D convolution networks (e.g., ResNeXt-101 [12]) are employed to extract motion information. We use convLSTM following behind the enhanced content feature maps to learn dynamic scene variations and object displacement transformations, and use LSTM in the motion stream to study consecutive motion changes. Finally, a multilayer perceptron (MLP) is exploited to fuse these features from two streams, leading to comprehensive video representations. Extensive experiments demonstrate that our method achieves state-of-the-art performance in the Vine benchmark dataset on Macro-F1 and Micro-F1. We further validate the effectiveness of our model in another dataset YUP++.

The main contributions of this work are threefold:

- We design a Global-Local Attention Module to learn discriminative semantic features from video content, where the global attention can capture scene-oriented information while the local attention can capture object-oriented information. It can be inserted to CNNs at any depth readily for further visual content representations.
- We propose a Hybrid-Attention Enhanced Two-Stream Fusion Network, which is capable of exploring and integrating scene-relevant, object-relevant and motion-relevant features to generate final comprehensive representations for the task of video venue category prediction.
- We conduct extensive experiments in the Vine benchmark dataset, demonstrating that the proposed model achieves state-of-the-art prediction performance on both Macro-F1 and Micro-F1. The visualization also shows the interpretability of our model.

II. Proposed Framework

Fig. 2 shows our proposed model HA-TSFN, which consists of two streams: the content stream and motion stream. The content stream mainly extracts appearance information. In this stream, we introduce a hybrid-attention network structure GLAM, and insert it to CNNs to generate discriminative visual content representations. In GLAM, the global attention learns contextual scene-oriented information via assigning channels with various weights while the local attention catches salient object-oriented features via allocating different weights for spatial regions. After crossing hybrid attentions, two types of feature maps are then merged by a convolution layer. The integrated feature maps generated by GLAM from multiple key frames are followed with convLSTM to learn long-term spatial-temporal content information. The motion stream jointly utilizes 3D neural networks and LSTM to catch consecutive movement variations. A MLP is finally used to fuse the features from two streams, with a softmax activation function followed to predict the probability of each class. In general, these two streams respectively describe different perspectives of the video, and both contribute to video venue category prediction. Next, we will detail the framework.

A. Content Stream

Each video has many frames and most of them are redundant and useless. Therefore, we extract key frames $V_Q = \{Q_1, ..., Q_t, ..., Q_L\}$ from each video, where Q_t represents the t-th key frame. Similar to [13], we extract key frames with equal intervals. The key frames captured by this method are relatively continuous, which benefits to sequential feature learning. In the content stream, different backbones can be used. Without loss of generality, we adopt ResNet-152 [11] as the backbone and utilize the output of the last convolution layer as content representations. The GLAM is introduced here to generate discriminative visual features, as illustrated in Fig. 2 with the red box. We then obtain expressive content features via convLSTM, which is used to capture long-term temporal dependencies such as dynamic scene variations and object location changes based on the output from the GLAM.

1) GLAM: GLAM consists of two attentions: a global attention and a local attention. The former is employed to collect scene-related information while the latter is used to capture object-related information. The GLAM combines global contextual knowledge with local salient regions via convolution layers for visual enhancement. We denote the content feature map $\boldsymbol{U}_t \in \mathbb{R}^{W\times H\times C}$ extracted from Q_t as the input of the GLAM, where W, H, C represent the width, height and channel dimension of feature maps, respectively.

Global Attention (GA). Due to limited receptive fields of kernels in CNNs, it is not feasible to utilize external information out of their receptive regions. In order to tackle this issue and make full use of contextual information to capture scene-oriented features, we design a global attention. Particularly, we apply a global average pooling operation on $\boldsymbol{U}_t$ to compress its spatial dimension and acquire a channel-wise vector $\boldsymbol{z}_t \in \mathbb{R}^C$. Here $z_{t,c}$ is calculated by $\boldsymbol{U}_{t,c}$, which is the c-th channel of $\boldsymbol{U}_t$:

$$z_{t,c} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \boldsymbol{U}_{t,c}(i,j) \tag{1}$$

where $\boldsymbol{z}_t$ represents the whole information of the c-th channel in $\boldsymbol{U}_t$, such as colors, shapes, and breaks the receptive field limitation. However, it is still unable to perceive other channels to capture global dependencies. In order to capture dependencies among channels, two MLPs are utilized to shrink and expand the vector $\boldsymbol{z}_t$:

$$\boldsymbol{a}_t^g = \sigma(\boldsymbol{W}_{expand}(\delta(\boldsymbol{W}_{shrink}\boldsymbol{z}_t))) \tag{2}$$

where $\boldsymbol{W}_{shrink} \in \mathbb{R}^{\frac{C}{r}\times C}$ are the weights of the shrinking layer to reduce the dimension of $\boldsymbol{z}_t$ and $\boldsymbol{W}_{expand} \in \mathbb{R}^{C\times \frac{C}{r}}$ are the weights of the expanding layer to map the vector into the original dimension space. σ is a sigmoid function and δ is a rectified linear function (ReLU). r is the compression ratio and $\boldsymbol{a}_t^g \in \mathbb{R}^C$ is the global attention score which represents the response of each channel. The designs of shrinking, expanding and compression ratio r are used to reduce model computation and keep generalization without lacking contextual information. Based on the global score, each channel can be distributed with a weight to realize reinforcement and the model is able to attend the whole information conveyed by content feature maps:

$$\boldsymbol{X}_t^g = \alpha(\tilde{\boldsymbol{A}}_t^g \circ \boldsymbol{U}_t) + \boldsymbol{U}_t \tag{3}$$

where $\tilde{\boldsymbol{A}}_t^g \in \mathbb{R}^{W\times H\times C}$ is generated by padding the vector $\boldsymbol{a}_t^g$ through its spatial dimension. $\circ$ denotes the Hadamard product. α is a hyperparameter that can adaptively balance the original input and contextual information. This attention can emphasize global information as well as retain the original knowledge.

Local Attention (LA). In order to discover more salient regions in key frames, we further introduce a local attention. Compared with the global attention which shrinks the content feature map $\boldsymbol{U}_t$ through its spatial dimension, the local attention squeezes $\boldsymbol{U}_t$ crossing its channel dimension:

$$\boldsymbol{A}_t^l = \rho(\boldsymbol{V}_{normal} * (\delta(\boldsymbol{V}_{shrink} * \boldsymbol{U}_t))) \tag{4}$$

where $\boldsymbol{V}_{shrink} \in \mathbb{R}^{1\times 1\times C\times \frac{C}{r}}$ are the kernels of the shrinking

convolution layer to compress the channel dimension of $\boldsymbol{U}_t$ and $\boldsymbol{V}_{normal} \in \mathbb{R}^{1\times1\times\frac{C}{r}\times1}$ are the kernels to normalize the channel dimension. δ is a ReLU function and ρ is a softmax function. $*$ denotes the convolution operator. $\boldsymbol{A}_t^l \in \mathbb{R}^{W\times H}$ represents the local attention score and it is applied on the content feature map $\boldsymbol{U}_t$:

$$\boldsymbol{X}_t^l = \beta(\tilde{\boldsymbol{A}}_t^l \circ \boldsymbol{U}_t) + \boldsymbol{U}_t \tag{5}$$

where $\tilde{\boldsymbol{A}}_t^l \in \mathbb{R}^{W\times H\times C}$ is produced by expanding $\boldsymbol{A}_t^l$ across its channel dimension. β is a hyperparameter that can selectively balance the original input and local salient regional features. Based on this attention, local information is highlighted and the original knowledge can also be reserved.

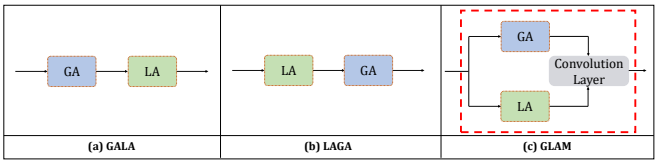


Fig. 3: Three different ways to arrange GA and LA. (a) GALA is a sequential way where GA is prior to LA. (b) LAGA is also a sequential way in which GA follows LA. (c) GLAM is a parallel way where GA and LA are arranged concurrently with a convolution layer followed for integration.

Convolutional Integration. $\boldsymbol{X}_t^g$ and $\boldsymbol{X}_t^l$ are highly complementary. The former is relevant to scene information and the latter is associated with object information. Therefore, we fuse them to generate comprehensive representations. In particular, we design three different ways to arrange the global and local attention. GALA is a sequential way where GA is prior to LA. LAGA is contrary to GALA. These two arrangement ways are showed in Fig. 3 (a) (b). Fig. 3 (c) permutes GA and LA parallelly with a convolution layer followed to integrate the outputs. Thus, two kinds of feature maps are concatenated in the corresponding channel c at the same position (i, j), generating the stacked data $\boldsymbol{X}_t^{cat}$:

$$\begin{cases} \boldsymbol{X}_{t,(i,j,1:c)}^{cat} = \boldsymbol{X}_{t,(i,j,1:c)}^{g} \\ \boldsymbol{X}_{t,(i,j,c:2c)}^{cat} = \boldsymbol{X}_{t,(i,j,1:c)}^{l} \end{cases} \tag{6}$$

where $\boldsymbol{X}_t^{cat} \in \mathbb{R}^{W\times H\times 2C}$ represents the concatenated feature maps. Subsequently, the output is passed into a convolution layer whose kernel size is 1×1 for fusion.

$$\boldsymbol{X}_t^e = \delta(\boldsymbol{f} * \boldsymbol{X}_t^{cat} + \boldsymbol{b}) \tag{7}$$

where $\boldsymbol{f} \in \mathbb{R}^{1\times1\times2C\times C}$ and $\boldsymbol{b} \in \mathbb{R}^C$ denote kernels and biases of the fusion layer, respectively. The convolution layer is used to reduce the channel dimensionality and realize weighted combinations of two feature maps [14]. Finally, the augmented feature map $\boldsymbol{X}_t^e$ is gained. In the final framework, we select convolutional integration to combine $\boldsymbol{X}_t^g$ and $\boldsymbol{X}_t^l$ for its excellent performance.

The design of GLAM is inspired by SE-Net [15] and the global attention in GLAM also refers to it, but with some differences as follows: 1) SE-Net only utilizes Squeeze-and-Excitation (SE) module to compute channel-wise weights for important information. Considering that each image contains not only contextual information but also salient regions, we add the local attention to the global attention with a convolutional integration in GLAM. The output of GLAM is more discriminative and can further enhance the information captured by the global attention through back propagation. 2) SE-Net regards the SE module as a separated block and it can be inserted into any depth of networks. However, we design GLAM as the independent module.

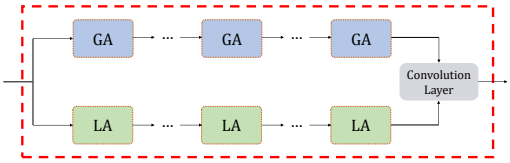


Fig. 4: The architecture of exGLAM. It contains N GAs and N LAs arranging parallelly with a convolution layer followed for integration.

2) Extending GLAM (exGLAM): In this subsection, we extend GLAM with multiple global and local attentions, as illustrated in Fig. 4. As mentioned before, the global attention can enhance scene-oriented features via catching contextual information, and the local attention can augment object-oriented features by capturing salient regions. Cascaded GA and LA can further respectively reinforce scene-oriented and object-oriented features to generate more discriminative visual representations. For simplicity, we set that the number of GA and LA are same, denoted as N. We define the GLAM with N GAs and N LAs arranging parallelly as exGLAM. It is described as follow:

$$\begin{cases} \boldsymbol{X}_t^g = \boldsymbol{GA}_N(\cdots \boldsymbol{GA}_k(\cdots \boldsymbol{GA}_1(\boldsymbol{U}_t))) \\ \boldsymbol{X}_t^l = \boldsymbol{LA}_N(\cdots \boldsymbol{LA}_k(\cdots \boldsymbol{LA}_1(\boldsymbol{U}_t))) \\ \boldsymbol{X}_t^e = \text{Convolutional Integration}(\boldsymbol{X}_t^{\boldsymbol{g}}, \boldsymbol{X}_t^{\boldsymbol{l}}) \end{cases} \tag{8}$$

where $\boldsymbol{GA}_k$ and $\boldsymbol{LA}_k$ represent the k-th global and local attention, respectively.

3) Sequence Learning: Sequential cues are vital for video venue prediction because they contain long-term spatial-temporal information such as dynamic scene variations and object location transformations. Here, we apply convLSTM [16] instead of LSTM in the content stream, because: (1) LSTM does not take spatial-temporal correlations into consideration. It inputs one-dimensional vectors, which collapses the spatial structure in feature maps as well as spatial-temporal relationships among input feature maps. In contrast, convLSTM that contains convolutional modules in the input-to-state and state-to-state transitions can retain spatial structures and learn temporal features. (2) There exists a lot of redundant information during the process of flatting feature maps to vectors and putting them into LSTM while convLSTM does not contain the information [16]. The augmented feature maps $\{\boldsymbol{X}_t^e\}_{t=1:L}$ are conveyed into convLSTM to learn long-term spatial-temporal representations:

$$\boldsymbol{X}_L^e = \text{convLSTM}(\{\boldsymbol{X}_t^e\}_{t=1:L}) \tag{9}$$

where $\boldsymbol{X}_L^e \in \mathbb{R}^{W\times H\times C}$.

Finally, we apply a spatial average pooling operation on $\boldsymbol{X}_L^e$ to obtain $\boldsymbol{x}^e \in \mathbb{R}^C$, which is the final representation of

the content stream.

B. Motion Stream

Motion information can also help recognize video venue categories for a video contains a series of movements that constitute a venue-related behavior. For example, in the Gym, people's movements are usually coherent and sequential, which is beneficial for inferring the venue category. Similar to [17], we preprocess the video into multiple clips defined as $V_P = \{P_1, ..., P_t, ..., P_{L'}\}$, where a sliding window manner is exploited to generate non-overlapped 16-frame clips. 3D convolution networks are performed on those clips to gain corresponding motion representations $\{\boldsymbol{x}_t^m\}_{t=1:L'}$, where P_t and $\boldsymbol{x}_t^m$ respectively represent the t-th clip and the extracted feature of the t-th clip. Here we employ a simple but efficient architecture ResNeXt-101 [12] as the backbone to extract motion features. We choose the output of the global average pooling layer as motion features because of their high-level semantics and low computation. As such, the acquired features contain spatial information and short-term temporal cues.

In order to learn long-term temporal dependencies among movements and catch a complete action of the video, the LSTM structure is applied. A list of motion features $\{\boldsymbol{x}_t^m\}_{t=1:L'}$ is delivered into LSTM to learn consecutive movements variants. We extract the last hidden representation $\boldsymbol{x}^m$ of LSTM as the motion stream output. In this stream, 3D CNNs are exploited to extract features which contain short-term spatial-temporal information. Under this condition, LSTM is able to learn long-term spatial-temporal motion information and reduce computation compared to convLSTM.

C. Two-Stream Fusion

We concatenate the enhanced content feature $\boldsymbol{x}^e$ and motion feature $\boldsymbol{x}^m$ to form a high-dimensional vector. The vector is delivered into a MLP to generate the final comprehensive representation. The cross entropy is adopted as the loss function.

III. Experiment

A. Experiment on Vine

We conducted our experiments in the large-scale benchmark dataset Vine [1]. The dataset contains 270,145 videos and each video is about 6 seconds long. The total number of categories is 188. Similar to [1], the dataset is split into three chunks: 80% of videos as the training set, 10% as the validation set and 10% as the testing set.

Implementation Details. Our model and all baselines are implemented by Tensorflow, and the Adam optimization is adopted. The framework adopts ResNet-152 pretrained on ImageNet dataset [11] as the backbone network in the content stream. We exploited its last convolution layer output to gain content feature maps and the dimension is set as: $W = 7$, $H = 7$ and $C = 2,048$. In the GLAM, the compression ratio $r = 4$ and the hyperparameters are initially set as $\alpha = \beta = 0$. Meanwhile, we employed the global average pooling layer output of ResNeXt-101 [12] pretrained on Kinetics [18] to obtain motion representations with 4,096 dimensions. A sliding window manner is used to generate input video clips where each video is divided into some non-overlapped 16-frame clips. The convLSTM and LSTM with 1,024 hidden sizes are applied in the content and motion stream, respectively. The kernel size of convLSTM is 1×1. The dropout ratio is 0.5 in convLSTM and LSTM. The last hidden state outputs of convLSTM and LSTM are final representations of two streams, which are then fused by a fully-connected layer with 1,024 dimensions. One key frame per second is selected to learn the content feature map. We set the fps (frames per second) as 16, and thus could get one clip per second for the motion feature. We set the batch size to 64, the epoch to 10 and the learning rate to 10^{-4} which is divided by 10 after 5 epochs.

Similar to [1], [3], Macro-F1 and Micro-F1 are adopted as metrics to evaluate the performance of our model. Macro-F1 gives equal weight to each class in the averaging process while Micro-F1 gives equal weight to each instance. When the values of Macro-F1 and Micro-F1 both achieve 1, the classification performance is the best.

Evaluation of Content Stream. In this subsection, we conducted experiments to show the effectiveness of each component in the content stream. Particularly, we first explored the importance of GLAM, then exGLAM and sequential information. Finally, we analyzed the effect of exGLAM with sequential information in the content stream. Note that we used a spatial average pooling operation to convert the 3D feature map to a one-dimensional vector and the default frame integration method is average pooling which is described as $\boldsymbol{v}^s = \frac{1}{|L|}\sum_{t=1:L} \boldsymbol{x}_t^s$, where $s = \{c, g, l, e\}$. Here $\boldsymbol{x}_t^c$, $\boldsymbol{x}_t^g$, $\boldsymbol{x}_t^l$ and $\boldsymbol{x}_t^e$ are acquired from the feature maps $\boldsymbol{U}_{t,c}$, $\boldsymbol{X}_t^g$, $\boldsymbol{X}_t^l$ and $\boldsymbol{X}_t^e$, respectively.

1) Evaluation of GLAM. We first compared global attention, local attention and their different arrangement ways in GLAM to verify their effectiveness. The baselines are listed as follow:

- **Content Feature MAP (CFM).** This baseline means that a list of feature maps from the 2D CNN is transformed into one-dimensional vectors via spatial average pooling, with a default frame integration method following to generate the final representation.
- **CFM-GA.** A series of feature maps are passed into the global attention to gain augmented feature maps, and a spatial average pooling operation is applied to generate one-dimensional vectors. The obtained vectors are fused to generate the final representation via the default frame integration method.
- **CFM-LA.** The baseline is similar to CFM-GA. However, it relies on the local attention to enhance feature maps.
- **CFM-GALA.** This baseline considers how to arrange two attentions to generate augmented feature maps. CFM-GALA is a sequential way where a list of feature maps first crosses the global attention, and then the local attention to gain enhanced 3D feature maps. The following operations are similar to CFM.
- **CFM-LAGA.** This baseline is also a sequential way, but a series of feature maps are first passed into the local

attention, and then the global attention.

- **CFM-gala [19].** It also contains two attentions. Specifically, the module exploits attention mechanism to generate a global score and a local score, and then fuses them to obtain the final weight for the input feature map. These weights are applied to original feature maps for enhancing visual representations.

Table I summarizes experimental results. From Table I, we could find: (1) CFM-GA respectively achieves 18.07% and 35.77% on Macro-F1 and Micro-F1, yielding 1.16% and 3.14% improvement compared to CFM, which indicates that the global attention can focus on different channels to enhance visual representations. (2) CFM-LA outperforms CFM with 2.0% and 0.69% margins on two metrics, respectively. This proves that the local attention can attend to salient regions for gaining more discriminative features. (3) CFM-GALA, CFM-LAGA, CFM-GLAM are all superior to CFM, which demonstrates that the features respectively captured by the global and local attention are complementary. Meanwhile, these three methods also exhibit different performance regarding different arrangement ways of GA and LA. CFM-GLAM gains the best performance with 26.48% on Macro-F1 and 40.70% on Micro-F1. This shows that the parallel arrangement of GA and LA can better encode the complementarity via convolution layers in the task of video venue category predication. (4) Our proposed CFM-GLAM surpasses the CFM-gala method with 8.97% and 5.23% margins on two metrics respectively, which proves that our design can better augment video visual representations by integrating two kinds of complementary feature maps than simply merging two scores.

Meanwhile, our design of the global attention refers to the SE module. In order to make our model more convincing, we conducted experiments to compare CFM-GLAM with SE-Net. The SE-Net achieves 20.38% on Macro-F1 and 35.21% on Micro-F1, which performs worse than CFM-GLAM. This result demonstrates that considering the local information and the integration method of different kinds of information can generate more discriminative representations.

TABLE I: Evaluations different types of attentions and their arrangements.

Method	Macro-F1	Micro-F1
CFM	16.91%	32.63%
CFM-GA	18.07%	35.77%
CFM-LA	18.91%	33.32%
CFM-GALA	20.23%	36.66%
CFM-LAGA	20.99%	37.06%
CFM-gala [19]	17.51%	35.47%
CFM-GLAM	**26.48%**	**40.70%**

In our proposed GLAM, we selected convolutional integration to fuse contextual information and salient regional features. We also conducted experiments on some other fusion methods to verify their effects. In Table II, we compared CFM-GLAM (Conv) with the following fusing methods: CFM-GLAM (Sum), CFM-GLAM (Max) and CFM-GLAM (Concat). The three baselines mean that using sum, max and concatenation fusion methods, respectively. From Table II, we could find that CFM-GLAM (Conv) achieves the best performance on both two metrics, which demonstrates that the convolutional integration can better encode the complementarity between two types of feature maps.

TABLE II: Evaluations of GLAM with different fusion methods.

Method	Macro-F1	Micro-F1
CFM-GLAM (Max)	18.29%	23.80%
CFM-GLAM (Sum)	20.12%	36.23%
CFM-GLAM (Concat)	20.94%	36.62%
CFM-GLAM (Conv)	**26.48%**	**40.70%**

In addition, we did some experiments on the effect of the hyperparameter r in GLAM. The results are shown in Table III, where $r = \{0, 1, 2, 4, 8, 16, 32\}$. Note that $r = 0$ means the CFM baseline that does not consider the attention mechanism. From Table III, we could see that when $r = 4$, Macro-F1 achieves the best and Micro-F1 reaches the best at $r = 1$. Considering the trade-off between performance and computation, we choose $r = 4$. Actually, using a uniform compression ratio may not be an optimal choice because different attentions have various roles. Further performance gains would be obtained after selecting the optimal compression ratio for each attention.

TABLE III: Evaluations of GLAM with different compression ratio.

r	Macro-F1	Micro-F1	Parameters	GFLOPs
0	16.91%	32.63%	0.4M	0.187
1	26.46%	**41.20%**	21.4M	494
2	26.21%	40.69%	15.1M	406
4	**26.48%**	40.70%	11.9M	361
8	26.42%	40.44%	10.3M	339
16	26.45%	40.55%	9.6M	328
32	26.42%	40.53%	9.2M	323

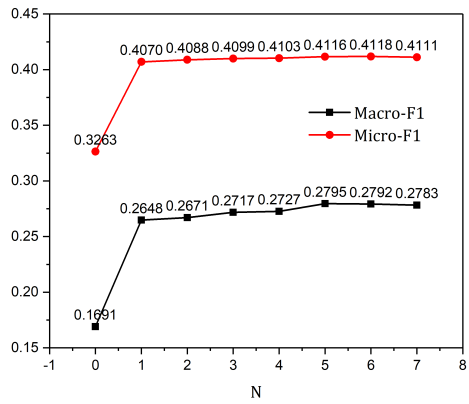


Fig. 5: Comparisons of CFM-exGLAM in the content stream with different numbers N of attentions where $N = 0$ means the CFM baseline and $N = 1$ means the CFM-GLAM baseline.

2) Evaluation of exGLAM. Fig. 5 compares CFM-exGLAM in the content stream with different numbers N of attentions where $N = \{0, 1, 2, 3, 4, 5, 6, 7\}$. From Fig. 5, it can be seen that there is a consistent promotion on both Macro-F1 and Micro-F1 when we increased N from 0 to 5. Then the performance begins to converge. When $N = 5$, the classification performance achieves the best. This phenomenon

shows that exGLAM can further enhance content feature maps via using multiple GAs and LAs. Multiple GAs can constantly augment the contextual information and multiple LAs can continually strengthen salient regional features, leading to more discriminative visual content representations.

3) Evaluation of Sequential Modules. We then evaluated the importance of sequential information in the content stream. Table IV shows experimental comparisons among average pooling, CFM-LSTM and CFM-convLSTM methods. CFM-LSTM utilizes LSTM to capture sequential information while CFM-convLSTM uses convLSTM. We could see that CFM-convLSTM achieves the best performance with 23.49% in Macro-F1 and 40.47% in Micro-F1, with 1.08% and 1.78% improvements on two metrics than CFM-LSTM. Compared with LSTM which destroys the spatial structure when it uses one-dimensional vectors as input, convLSTM can retain spatial structures and better encode long-term spatial-temporal information, leading to expressive feature representations.

TABLE IV: Evaluation of sequential modules in the content stream.

Method	Macro-F1	Micro-F1
CFM	16.91%	32.63%
CFM-LSTM	22.41%	38.69%
CFM-convLSTM	**23.49%**	**40.47%**

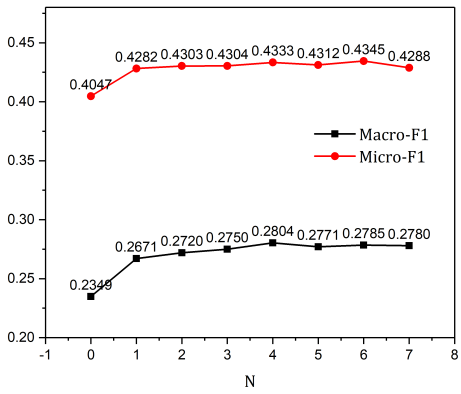


Fig. 6: Comparisons of CFM-exGLAM-convLSTM in the content stream with different numbers N of attentions, where $N = 0$ means the CFM-convLSTM baseline and $N = 1$ means the CFM-GLAM-convLSTM baseline.

4) Evaluation on the combination of exGLAM and convLSTM. We evaluated the combination of exGLAM and convLSTM in the content stream as illustrated in Fig. 6. From Fig. 6, we could see (1) The method CFM-GLAM-convLSTM surpasses CFM-GLAM and CFM-convLSTM, yielding 26.71% and 42.82% on two metrics respectively. This result shows that the combination of GLAM and convLSTM can further improve prediction performance compared to the single component. (2) The method CFM-exGLAM-convLSTM achieves the best performance with 28.04% on Macro-F1 and 43.33% on Micro-F1 when $N = 4$. This demonstrates that exGLAM with convLSTM can generate more discriminative spatial-temporal features via multiple attentions than GLAM with convLSTM. (3) Compared to CFM-exGLAM where $N = 5$ is the peak point, CFM-exGLAM-convLSTM performs best at $N = 4$ and its performance improvement is lower than that of the former. The probable reason is that introducing sequential information may influence the effect of exGLAM and make the performance converge more quickly. (4) When $N > 4$, CFM-exGLAM-convLSTM is slightly inferior to CFM-exGLAM on Macro-F1. We inferred that using multiple attentions and convLSTM brings too many parameters, leading to model overfitting and performance dropping. The trend on Micro-F1 when $N > 4$ also confirms our conjecture.

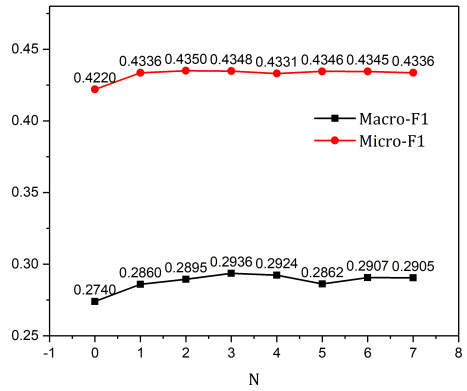


Fig. 7: Comparisons of exGLAM in our framework with different numbers N of attentions.

Comparisons of exGLAM with different numbers of attentions in our framework. Fig. 7 compares exGLAM in our framework with different numbers N of attentions where $N = \{0, 1, 2, 3, 4, 5, 6, 7\}$. When $N = 0$, we only considered sequential information and motion stream without attention mechanism. Our model achieves the best performance with 29.36% and 43.48% on Macro-F1 and Micro-F1 respectively when $N = 3$, which proves that attentions, sequential information and motion stream all contribute to video venue prediction. The performance reaches saturation and drops a little when $N > 3$, resulting from the fact that our model overfits for introducing too many parameters. We could also see that the peak point moves forward compared to CFM-exGLAM-convLSTM. This phenomenon is caused by introducing motion stream in our framework, which may accelerate the convergence process.

TABLE V: Comparisons with state-of-the-art methods in Vine.

Method	Macro-F1	Micro-F1
TRUMANN (AlexNet) [1]	5.21%	25.27%
HCM-FL (VGG16) [3]	18.82%	37.40%
DARE (ResNet-50) [2]	25.83%	39.59%
NMCL (ResNet-50) [20]	27.42%	41.11%
HA-TSFN (VGG16, GLAM)	22.04%	39.05%
HA-TSFN (ResNet-50, GLAM)	28.25%	42.78%
HA-TSFN (ResNet-152, GLAM)	28.60%	43.36%
HA-TSFN (VGG16, exGLAM)	22.63%	39.32%
HA-TSFN (ResNet-50, exGLAM)	28.41%	42.89%
HA-TSFN (ResNet-152, exGLAM)	**29.36%**	**43.48%**

Comparisons with state-of-the-art methods. We compared our model against state-of-the-art methods on video venue prediction in Table V. TURMANN, DARE and NMCL all use additional modalities (e.g., textual and acoustic) to

produce final representations while TURMANN, DARE and HCM-FL introduce external datasets or hierarchy venue category structure for feature learning. In our model, we only considered visual features without other modalities or datasets. For fair comparisons, we conducted experiments with different backbones in the content stream, such as VGG16 [21] and ResNet-50 [11]. For simplicity, we still used ResNeXt-101 to extract motion features. From Table V, we could see that HA-TSFN surpasses the state-of-the-art methods with the same backbone. Meanwhile, as the backbone gradually improves, the classification performance increases. We utilized exGLAM in different backbones to further prove its effectiveness. In exGLAM, we set $N = 3$ for its excellent performance. By introducing exGLAM, all backbone networks perform better on two metrics than using GLAM. HA-TSFN with ResNet-152 and exGLAM achieves the best performance with 29.36% and 43.48% on Macro-F1 and Micro-F1, respectively.

Complexity Analysis. We analyzed the computational cost and parameters on our proposed method HA-TSFN and other baselines, such as DARE [2] and NMCL [20]. In order to discover which module has the highest computation, we also conducted evaluations on the attention mechanism, sequential modules and motion stream. The experimental results are shown in Table VI. From Table VI, we could find that 1) The results among CFM, CFM-GA and CFM-LA indicate that attention mechanism brings the computational increase for it performs on multiply feature maps to realize enhancement. 2) The baseline CFM-GLAM adopts the convolutional integration method, which introduces higher computation and more parameters. However, this fusion method is optimal and achieves the best performance. 3) The convLSTM also leads to additional computation for it contains convolutional modules in the input-to-state and state-to-state transitions. 4) It is obvious that the motion stream improves performance and raises computation simultaneously in our proposed method. 5) Our model HA-TSFN has the best classification accuracy with higher computation compared to other methods for the combination of attention mechanism, convLSTM and motion stream.

TABLE VI: Comparisons with existing methods on computation and parameters.

Method	Parameters	FLOPs
CFM	0.4M	0.19G
CFM-GA	2.5M	5.38G
CFM-LA	1.4M	40.00G
CFM-GLAM	11.9M	361.36G
CFM-convLSTM	12.8M	79.11G
CFM-GLAM-convLSTM	24.3M	440.25G
HA-TSFN (GLAM)	39.0M	445.93G
DARE	7.1M	14.25M
NCML	27.3M	82.3M

In summary, we studied the effectiveness of attention mechanism, sequential modules and motion stream in our proposed model. In the Content Stream, we first explored the influence of GLAM, including the types of attentions, the number of attentions, the arrangement ways of attentions and the fusion methods of attentions. Second, we investigated which sequential modules could learn better spatial-temporal features. Meanwhile, we conducted experiments on the combination of attention mechanism and sequential modules with different number of attentions to show their collaboration. Third, we did some experiments to demonstrate the effect of the motion stream. Then we researched the impact of different number of attentions in the final model and used the best performance to compare with other related works. Finally, we analyzed the complexity and parameters of our model.

Classification Histogram. We illustrated prediction results of some baselines and our model HA-TSFN on 25 venue categories to better analyze the performance. From Fig. 8, we could observe that (1) Our model outperforms baselines on most categories, which indicates that our design is effective and promising. (2) The methods CFM-GLAM, CFM-exGLAM and CFM-convLSTM are all superior to CFM. This phenomenon shows that attention mechanism and sequential cues can influence prediction performance positively. (3) The method CFM-GLAM-convLSTM has a little performance drop compared to HA-TSFN (GLAM) on most categories, which proves the effectiveness of the motion stream. (4) HA-TSFN (exGLAM) exceeds HA-TSFN (GLAM), which demonstrates that exGLAM with multiple attentions can generate more discriminative visual features.

Visualization. In Fig. 9, four video samples from different categories are presented. We visualized the feature maps crossing attentions via Gradient-weighted Class Activation Heatmap on Image (Grad-CAM). Considering that convolutional integration, temporal dependencies and motion information may influence the visualization, we drew heatmaps excluding them to better visualize.

From the visualization, we could see that (1) After crossing the global attention, feature maps tend to care about contextual information, such as scene-related regions. For example, in Fig. 9 (a), the global attention captures all instances and their position layouts, which constitute the Basketball Stadium scene via assigning channels with various weights. (2) When the local attention is applied, feature maps tend to focus more on salient areas, such as certain objects. As illustrated in Fig. 9 (a), the local attention mainly focuses on the athletes and basketball via allocating different weights for spatial regions. It is clear that attention mechanism can adaptively catch the information related to the ground-truth class and emphasize them. Fig. 9 (b) is another sample of the Beach venue. The global attention attends to the scene that consists of the beach, trees, persons, road and their arrangements while the local attention only notices the beach and trees. From qualitative results, we could discover that the global attention focuses on scene-oriented information and the local attention cares about object-oriented regions, which can enhance video semantic representations and improve the interpretability of our model.

TABLE VII: Quantitative evaluations on the quality of attention maps.

Data	CFM	CFM-GA	CFM-LA	CFM-GLAM	Gound Turth
Sample (a)	63.02%	68.42%	66.47%	**98.10%**	Basketball Stadium
Sample (b)	26.03%	59.47%	63.39%	**92.03%**	Beach
Sample (c)	31.15%	33.12%	35.85%	**89.28%**	Garden
Sample (d)	59.99%	79.40%	70.89%	**85.38%**	Football Stadium
Test Data	32.63%	35.77%	33.32%	**40.70%**	-

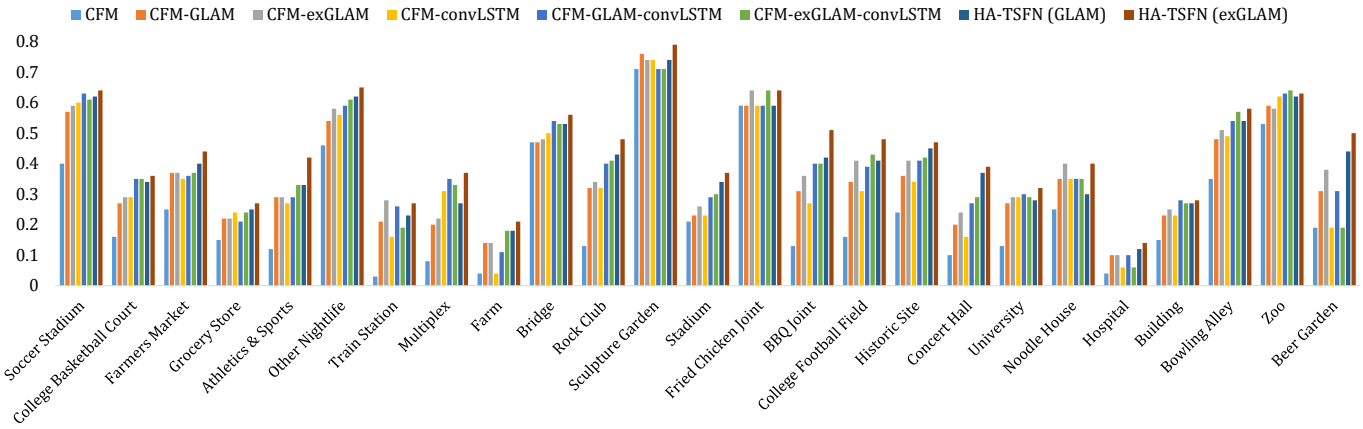


Fig. 8: Video venue prediction results of some baselines and our model HA-TSFN on 25 categories.

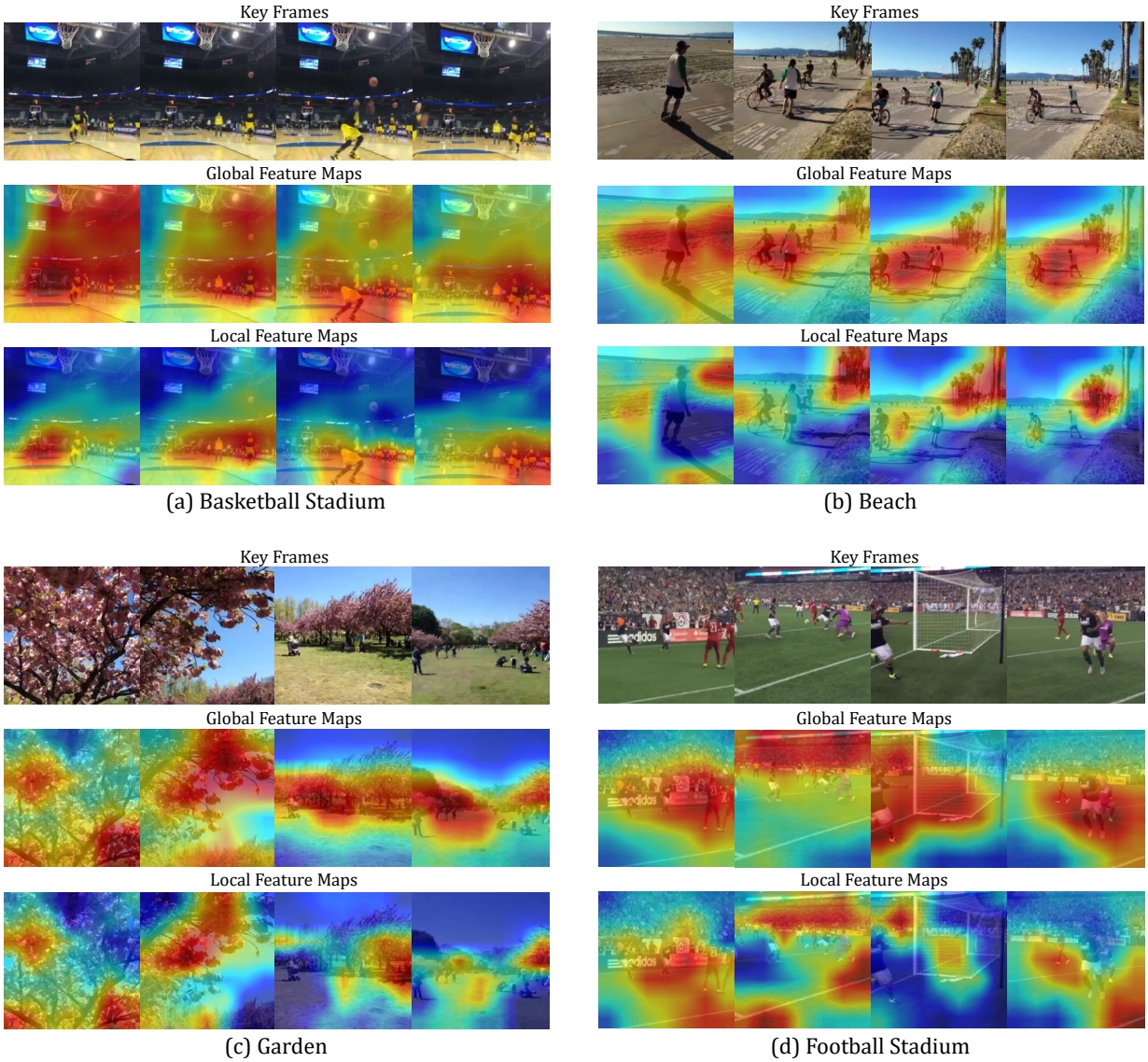


Fig. 9: Qualitative results. We selected some video samples from the Vine dataset to show where and what the attention mechanism focuses on. We illustrated the key frames extracted from videos and the feature maps crossing the global and local attention, respectively. The red color represents salient and important regions and the blue color is on the contrary.

TABLE VIII: The ablation studies on stationary and moving dataset in YUP++.

Content Stream				Motion Stream	Stationary		Moving	
CFM	GLAM	exGLAM	convLSTM	LSTM	Macro-F1	Micro-F1	Macro-F1	Micro-F1
✓					89.62%	90.00%	80.58%	80.96%
✓	✓				90.57%	90.74%	81.09%	81.48%
✓		✓			91.18%	91.30%	83.31%	83.70%
✓			✓		91.33%	91.48%	81.64%	82.22%
✓	✓		✓		91.73%	91.85%	83.49%	83.52%
✓		✓	✓		92.15%	92.22%	84.47%	84.81%
✓	✓		✓	✓	93.83%	93.89%	85.48%	85.74%
✓		✓	✓	✓	**95.32%**	**95.37%**	**86.68%**	**86.85%**

We also showed the quality of attention maps quantitatively. Inspired by [22], [23], we computed the softmax activation scores of the ground truth class at the video-level. The samples in Table VII are from Fig. 9. In order to make our quantitative analysis comprehensive, we showed the overall performance of the testing data. We utilized the Micro-F1 to evaluate the quality of these attention maps produced by the methods CFM, CFM-GA, CFM-LA and CFM-GLAM. From the results in Table VII, we could find that the attentions can enhance feature maps and the GLAM reaches the best reinforcement effect.

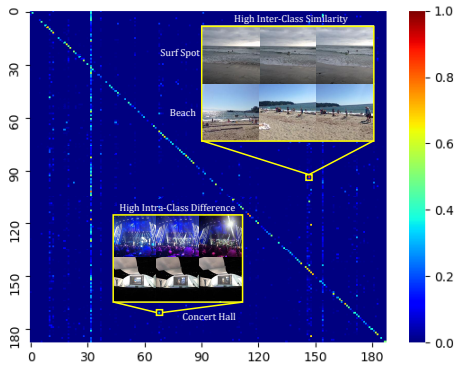


Fig. 10: The confusion matrix of our model HA-TSFN on Vine dataset. The horizontal axis means prediction classes and the vertical axis means ground-truth classes. The diagonal indicates the predicrion performance of our model.

Confusion Matrix. We showed the confusion matrix of our model HA-TSFN to better analyze the performance. From the matrix, we could see that our model can predict correctly with a high probability on certain categories. For example, our model achieves 66.67% classification accuracy on the Art Museum class. However for some other categories, the performance is comparatively inferior. The reasons are as follows: (1) High inter-class similarity. For example, the class Beach and Surf Spot are so similar that even humans can not identify them correctly. (2) High intra-class difference. As illustrated in Fig. 10, the class Concert Hall contains two samples which vary largely. And (3) the low quality of some categories. There is a clear vertical line on the City class in the confusion matrix, which means that some samples from other classes are predicted as City. Further observation tells us that the granularity of the City category is coarse, resulting in other messy videos falling into this class.

B. Experiment on YUP++

We further conducted evaluation in another dataset YUP++ to verify the effectiveness of our model. The YUP++ dataset [24] is composed of 1,200 videos with 20 scene classes. Half of videos are obtained by a static camera and the other half are caught by a moving camera which includes pan, tile, zoom and jitter. The average length of a video is about 5 seconds. The train/test ratio is 10/90. In this dataset, we set the fps as 32 to gain motion features and selected 4 key frames per second to acquire content feature maps. We set the batch size to 15, the epoch to 20, the learning rate to 10^{-4} which is divided by 10 after 10 epochs. For exGLAM, we selected different N to make all methods achieve the best performance. The other configurations are the same as Vine.

TABLE IX: Comparisons with state-of-the-art methods in YUP++.

Methods	Stationary	Moving
TRN (ResNet-50) [24]	92.40%	81.50%
SVMP [25]	92.50%	83.10%
GRP [26]	92.90%	83.60%
DSP (Inception-ResNet-v2) [27]	95.10%	88.30%
HAF+Bow/FV halluc (I3D) [28]	94.81%	**89.63%**
MSOE-two-stream (C3D) [29]	**97.00%**	87.00%
HA-TSFN (VGG16, GLAM)	90.56%	79.63%
HA-TSFN (ResNet-50, GLAM)	93.70%	83.52%
HA-TSFN (ResNet-152, GLAM)	93.89%	85.74%
HA-TSFN (VGG16, exGLAM)	91.11%	80.74%
HA-TSFN (ResNet-50, exGLAM)	94.63%	85.00%
HA-TSFN (ResNet-152, exGLAM)	95.37%	86.85%

We did ablation studies on the stationary and moving dataset in YUP++. Table VIII presents experimental results. The performance shows that each component and the combinations of different components have positive impacts on improving prediction performance, which demonstrates that attention mechanism, sequential information and motion stream all contribute to generating discriminative feature representations. Table IX compares our model against start-of-the-art methods. In general, our model outperforms TRN, SVMP and GRP in both stationary and moving datasets. However our model does not exceed MSOE-two-stream in the stationary dataset as well as DSP, HAF+Bow/FV halluc and MSOE-two-stream in the moving dataset. We analyzed the reason is that compared to our model which only utilizes visual information, these models all harness external knowledge, such as adversarial noises, improved dense trajectory descriptors, optical flows and additional datasets to enhance their classification abilities. We also investigated the performance of our model with different

backbones in Table IX. We could see that the performance improves with more advanced backbones used. Note that SVMP and GRP do not claim their backbones.

Compared the experimental results in Table V and Table IX, we could discover that the model with exGLAM improves slightly over GLAM in the Vine. However, exGLAM increases 1.0%-1.5% in the YUP++, which demonstrates its effectiveness. We analyzed the reasons from two perspectives, datasets and model. The different characteristic of datasets is the main reason to lead this phenomenon. Vine is collected from social media platforms which contains more complicated scenes and more objects than YUP++, bringing hard recognition. And the videos in YUP++ are shot by professional cameras rather than mobile phones, improving the quality of the dataset. The other reason we inferred is that utilizing sequential information and motion stream increase parameters in the model. Meanwhile, exGLAM with multiple GAs and LAs also brings external parameters. Plenty of parameters make the model prone to overfitting, resulting in slight improvement of exGLAM.

IV. Related Work

Our work is closely related to the following fields: 1) venue category prediction, 2) long-term temporal dependency learning, 3) attention learning and 4) two-stream network.

A. Venue Category Prediction

Many researches have studied venue prediction or location recognition from images [6], [30] or videos [1], [31]. Venue prediction belongs to the POI-level location recognition which is important in the real world, especially for the social media network. Many researchers are dedicated to tackling this problem. For example, Chen *et al.* [30] discovered business-aware visual concepts and realized multimodal business venue recognition from images. In [32], a multi-scale atrous convolution network was proposed for food-related place classification based on egocentric photos.

In addition to image venue prediction, more and more studies focus on venue category prediction from videos. For example, Zhang *et al.* [1] first introduced the task and built a tree-guided multi-task multi-modal learning model to estimate venue categories. Nie *et al.* [2] improved classification performance by harnessing external sound knowledge to enhance acoustic modality. Jiang *et al.* [3] introduced a hierarchy-dependent cross-platform multi-view feature learning framework. They used cross-platform data as additional information to reinforce visual features. Considering the complementarity and consistency among multiple modalities, Wei *et al.* [20] presented a multimodel cooperative learning model to incorporate these information explicitly. Due that videos especially from social websites have few textual descriptions and much noisy audio information, we focused on visual reinforcement for the task of video venue prediction.

B. Long-Term Temporal Dependency Learning

Commonly used models dealing with video data, such as C3D [17] and I3D [18], cannot extract long-term temporal dependencies in videos. Karpathy *et al.* [33] proposed three methods to realize long-term temporal features fusion: early fusion, slow fusion and late fusion. However, the experimental results show that these fusion methods could not improve performance significantly. Feichtenhofer *et al.* [14] proposed 3D convolution fusion and 3D pooling for local spatial-temporal features to get long-term sequential cues. LSTM [34] is one of the most popular networks on learning long-term temporal dependencies because it can solve the gradient vanishing problem. Donahue *et al.* [35] designed a recurrent convolution network followed by LSTM to capture long-term spatial-temporal features for activity recognition, image caption and video description. Nevertheless, the inputs of the LSTM must be one-dimensional vectors which collapse spatial structures in videos. The convLSTM is then proposed to maintain spatial structures and learn long-term spatial-temporal relationships among feature maps as well as filter redundant information [16]. Motivated by them, we utilized both LSTM and convLSTM in our framework to capture long-term spatial-temporal information.

C. Attention Learning

Since attention mechanism has been first proposed in [36], it becomes more and more popular because of its superior performance [37]–[39]. The attention mechanism can filter redundant information to make the network focus more on important regions. It is mainly divided into two types: hard attention and soft attention. Hard attention always makes binary choices and it is not differentiable during the training process. Some works such as [40] and [41] exploit hard attention to select a series of key regions for object recognition. In contrast, soft attention uses weighted average method instead of hard choices. Based on it, Linsley *et al.* [19] introduced a global-and-local (gala) module to combine local saliency with global contextual information for generating discriminative representations on the task of image classification. They gained a global score and a local score via gala, and then fused them to produce the final weight. Li *et al.* [42] proposed a multi-level attention model with a frame-level attention and a region-level attention. They both captured visual features, however the region-level attention cares about salient regions in each frame and the frame-level attention focuses on the relationship among frames. Different from them, in our model, we designed the GLAM which consists of a global attention to catch scene-oriented features, a local attention to catch object-oriented features and a convolution layer to integrate these information.

D. Two-Stream Network

The two-stream network is very popular in many fields, such as video classification [10], [14], [17], [22] and object tracking [43]–[46]. In [10], Simonyan *et al.* firstly proposed a two-stream convolutional network which incorporated RGB and optical flow streams for video action classification. Feichtenhofer *et al.* [14] explored different ways to integrate two streams spatially and temporally for better feature representations. The appearance of [17] made a transition from 2D

CNNs to 3D CNNs, which can learn better spatial-temporal information. However, 3D CNNs are prone to overfitting in small datasets. In order to deal with this problem, Kinetics [18] was introduced as a large scale video dataset to train 3D CNNs successfully. Inspired by them, we utilized the two-stream network with an ImageNet-pretrained 2D CNN to extract content features and a Kinetics-pretrained 3D CNN to extract motion features.

V. Conclusions

In this paper, we propose a Hybrid-Attention Enhanced Two-Stream Fusion Network for the video venue prediction task. First, we design a GLAM in the content stream to learn more discriminative visual representations, where a global attention is utilized to catch scene-oriented features via assigning channels with various weights while a local attention is employed to capture object-oriented features via allocating different weights for spatial regions. We also extend GLAM with multiple attentions, namely exGLAM to further explore the effectiveness of attention mechanism. Second, considering sequential cues in video content, convLSTM is used in the content stream to learn long-term spatial-temporal features. Then, we combine the content stream with motion information to generate final comprehensive representations, forming the HA-TSFN framework. Finally, we conduct qualitative and quantitative experiments in the Vine benchmark dataset, and the experimental results demonstrate the effectiveness of our model. Furthermore, the attention mechanism in GLAM increases the interpretability of our model.

Our model also has some limitations which can guide our future work: 1) Our model is weak in computation. It takes feature maps as input and introduces the GLAM module including a convolution layer, increasing the computation. Considering that, we plan to design a lightweight GLAM, which can reduce the computation without affecting performance and interpretability. 2) We utilize a MLP to integrate the content stream and motion stream without considering richer relationships between them. Based on it, we expect to capture more relationships, such as conflict and consistency, to generate discriminative features.

References

[1] J. Zhang, L. Nie, X. Wang, X. He, X. Huang, and T. S. Chua, "Shorter-is-better: Venue category estimation from micro-video," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1415–1424.

[2] L. Nie, X. Wang, J. Zhang, X. He, H. Zhang, R. Hong, and Q. Tian, "Enhancing micro-video understanding by harnessing external sounds," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1192–1200.

[3] S. Jiang, W. Min, and S. Mei, "Hierarchy-dependent cross-platform multi-view feature learning for venue category prediction," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1609–1619, 2018.

[4] Y. Zhu, J. Wang, L. Xie, and L. Zheng, "Attention-based pyramid aggregation network for visual place recognition," in *Proceedings of the 26th ACM International Conference on Multimedia*, 2018, pp. 99–107.

[5] T. Weyand, I. Kostrikov, and J. Philbin, "Planet-photo geolocation with convolutional neural networks," in *European Conference on Computer Vision*, 2016, pp. 37–55.

[6] N. Vo, N. Jacobs, and J. Hays, "Revisiting im2gps in the deep learning era," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2621–2630.

[7] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1437–1445.

[8] M. M. K. Sarker, H. A. Rashwan, F. Akram, E. Talavera, S. F. Banu, P. Radeva, and D. Puig, "Recognizing food places in egocentric photo-streams using multi-scale atrous convolutional networks and self-attention mechanism," *IEEE Access*, vol. 7, pp. 39 069–39 082, 2019.

[9] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems 29*, 2016, pp. 892–900.

[10] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 568–576.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[12] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[13] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2630–2640.

[14] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[16] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, 2015, pp. 802–810.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[18] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[19] D. Linsley, D. Schiebler, S. Eberhardt, and T. Serre, "Learning what and where to attend," in *International Conference on Learning Representations*, 2019.

[20] Y. Wei, X. Wang, W. Guan, L. Nie, Z. Lin, and B. Chen, "Neural multimodal cooperative learning toward micro-video understanding," *IEEE Transactions on Image Processing*, vol. 29, pp. 1–14, 2019.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[22] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "Mars: Motion-augmented RGB stream for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7882–7891.

[23] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[24] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Temporal residual networks for dynamic scene recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4728–4737.

[25] J. Wang, A. Cherian, F. Porikli, and S. Gould, "Video representation learning using discriminative pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1149–1158.

[26] A. Cherian, B. Fernando, M. Harandi, and S. Gould, "Generalized rank pooling for activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3222–3231.

[27] J. Wang and A. Cherian, "Learning discriminative video representations using adversarial perturbations," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 685–701.

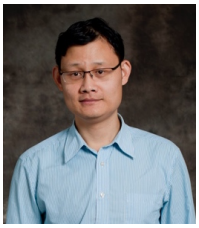
[28] L. Wang, P. Koniusz, and D. Q. Huynh, "Hallucinating IDT descriptors and i3d optical flow features for action recognition with cnns," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8698–8708.

[29] I. Hadji and R. P. Wildes, "A new large scale dynamic texture dataset with application to convnet understanding," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 320–335.
[30] B.-C. Chen, Y.-Y. Chen, F. Chen, and D. Joshi, "Business-aware visual concept discovery from social media for multimodal business venue recognition," in *Thirtieth AAAI Conference on Artificial Intelligence*, pp. 61–68.
[31] G. Friedland, J. Choi, H. Lei, and A. Janin, "Multimodal location estimation on flickr videos," in *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, 2011, pp. 23–28.
[32] M. K. Sarker, H. A. Rashwan, E. Talavera, S. Furruka Banu, P. Radeva, D. Puig *et al.*, "MACNet: Multi-scale atrous convolution networks for food places classification in egocentric photo-streams," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 0–0.
[33] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.
[34] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," in *Advances in Neural Information Processing Systems*, 1997, pp. 473–479.
[35] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.
[36] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.
[37] Q. Lai, W. Wang, H. Sun, and J. Shen, "Video saliency prediction using spatiotemporal residual attentive networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1113–1126, 2019.
[38] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 1531–1544, 2018.
[39] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, pp. 38–49, 2017.
[40] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems*, 2014, pp. 2204–2212.
[41] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," in *International Conference on Learning Representations*, 2015.
[42] X. Li, B. Zhao, and X. Lu, "MAM-RNN: Multi-level attention model based rnn for video captioning," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 2208–2214.
[43] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 459–474.
[44] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, and F. Porikli, "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Transactions on Image Processing*, vol. 28, pp. 3516–3527, 2019.
[45] Z. Liang and J. Shen, "Local semantic siamese networks for fast tracking," *IEEE Transactions on Image Processing*, vol. 29, pp. 3351–3364, 2019.
[46] J. Shen, X. Tang, X. Dong, and L. Shao, "Visual object tracking by hierarchical attention siamese network," *IEEE transactions on cybernetics*, vol. 50, pp. 3068–3080, 2019.

Yanchao Zhang received the B.E. degree in the School of Computer Science and Technology, Shandong University, Jinan, China. She is currently pursuing the M.E. degree with the School of Computer Science and Technology, Shandong University, Tsingtao , China. Her research interests are multimedia computing and compution vision.

Weiqing Min received the B.E. degree from Shandong Normal University, Jinan, China, in 2008 and M.E. degree from Wuhan University, Wuhan, China, in 2010, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2015, respectively. He is currently an associate professor at the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences. His current research interests include multimedia content analysis, understanding and applications, food computing and geo-multimedia computing. He has authored and co-authored more than 40 peer-referenced papers in relevant journals and conferences, including ACM Computing Surveys, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, ACM TOMM, IEEE Multimedia Magazine, ACM Multimedia, AAAI, IJCAI, etc. As the lead guest editor, he organized special issues on international journals, including IEEE Multimedia Magazine, Multimedia Tools and Applications. He has served as one TPC member of many academic conferences, including ACM MM, AAAI, IJCAI, etc. He is the recipient of 2016 ACM TOMM Nicolas D. Georganas Best Paper Award and the 2017 IEEE Multimedia Magazine Best Paper Award.

Liqiang Nie is currently a professor with Shandong University. Meanwhile, he is the dean with the Shandong AI institute. He received his B.Eng. and Ph.D. degree from Xi'an Jiaotong University in July 2009 and National University of Singapore (NUS) in 2013, respectively. After PhD, Dr. Nie continued his research in NUS as a research follow for three years. His research interests lie primarily in multimedia computing and information retrieval. Dr. Nie has co-/authored more than 150 papers and 4 books, received more than 7,700 Google Scholar citations as of Aug 2020. He is an AE of Information Science, and an area chair of ACM MM 2018-2020. He has received many awards, like sigir best paper honorable mention in 2019 and sigmm rising star in 2020.



Shuqiang Jiang (SM'08) is a professor with the Institute of Computing Technology, Chinese Academy of Sciences(CAS), Beijing and a professor in University of CAS. He is also with the Key Laboratory of Intelligent Information Processing, CAS. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 150 papers on the related research topics. He was supported by the New-Star program of Science and Technology of Beijing Metropolis in 2008, NSFC Excellent Young Scientists Fund in 2013, Young top-notch talent of Ten Thousand Talent Program in 2014. He won the Lu Jiaxi Young Talent Award from Chinese Academy of Sciences in 2012, and the CCF Award of Science and Technology in 2012. He is the senior member of IEEE and CCF, member of ACM, Associate Editor of ACM TOMM, IEEE Multimedia, Multimedia Tools and Applications. He is the vice chair of IEEE CASS Beijing Chapter, vice chair of ACM SIGMM China chapter. He is the general chair of ICIMCS 2015, program chair of ACM Multimedia Asia2019 and PCM2017. He has also served as a TPC member for more than 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, IJCAI, AAAI, ICME, ICIP, and PCM.