Illustration of Video Visual Relation Detection Based on Graph Neural Network

MINGCHENG QU$^{1,2}$, JIANXUN CUI$^{1,2}$, YUXI NIE$^{1}$, AND TONGHUA SU$^{1,2}$

$^{1}$Department of Software, Harbin Institute of Technology, Harbin 150040, China
$^{2}$Chongqing Research Institute of HIT, Chongqing 400000, China

Corresponding author: Yuxi Nie (yuxi.nie@foxmail.com)

This work was supported by the National Natural Science Foundation of China under Grant 61402131.

Abstract

Visual relation detection task is the bridge between semantic text and image information, and it can better express the content of images or video through the relation triple <subject, predicate, object>. This significant research can be applied to image question answering, video subtitles and other directions. Using the video as input to the task of visual relationship detection receives less attention. Therefore, we propose an algorithm based on the graph convolution neural network and multi-hypothesis tree to implement video relationship prediction. Video visual relationship detection algorithm is divided into three steps: Firstly, the motion trajectories of the subject and object in the input video clip are generated; Secondly, a VRGE network module based on the graph convolution neural network is proposed to predict the relationship between objects in the video clip; Finally, the relationship triplets are formed through the multi-hypothesis fusion algorithm (MHF) and the visual relationship. We have verified our method on the benchmark ImageNet-VidVRD dataset. The experimental results demonstrate that our proposed method can achieve a satisfactory accuracy of 29.05% and recall of 10.18% for visual relation detection.

Index Terms

Video visual relation detection, target detection, graph convolutional neural network.

I. INTRODUCTION

Visual Relation Detection (VRD) is a more complex task than Object Detection, in addition to obtain object category and the position, it needs to use <subject, predicate, object> to describe the relationship between two objects. Both subject and object are category labels of detected targets, and predicates are not only limited to the scope of verbs, but also prepositional phrases expressing spatial relations, etc. In this process, the input dataset is partially annotated, and further detailed annotation is realized through deep learning neural network in the following operations [1]. By outputting a large number of relational triplets, the interpretation of images and the construction of semantic information are completed. Due to the types of objects and expressions of the connection are varied, there are thousands of combinations of relational triplets, which need to be screened in the detection process to get the most appropriate expression.

According to the different ways of processing objects, using the video as the input is more challenging [2]. In current research tasks, the relationship detection for static images is more than that for dynamic video. Although the relationship between objects in video is an important factor for in-depth understanding of dynamic visual content, there are few researches on video-level relationship detection and reasoning because video-level relationship detection is much more difficult than image-level relationship detection [3]. The relational triplets obtained by each fragment need to be fused. In addition, the relationship is no longer just static in the detection process, it also has the dynamic expression, for example, “faster than”, “run to”, “on the left”, and so on. The dynamic relationship can only be detected by analyzing video clips in continuous time series. Target tracking technology is also used in visual relationship detection at the video level, which is different from target detection. In the process of tracking, the target object should be separated from the background at all times. In addition to marking the position of the target, the whole moving process should be recorded. These all increase a lot of computation, so video-level visual relationship detection is more complex and workload.

There are several significances to study this topic. First, the study of visual relation detection can help to realize more abstract semantic understanding of images; Second, the study of video dynamic relations can help computers to
build more efficient models in complex visual tasks, such as visual question-and-answer, video retrieval and so on; Third, providing a unified visual relationship detection framework, so as to faster identify the relationship between objects, such as comparative relationship, coordinate connection and behavior; Fourth, apply the results of visual relationship detection at the video level to real life, such as assisting police officers to track targets through the monitoring system. With the development of computer hardware and the update of deep learning field models, machines will have a deeper understanding of images and the results will have more practical application value.

The goal of this paper is to detect the visual relationship of objects in the video, classify the objects in the relationship triple more accurately, obtain predicate prediction with high confidence as far as possible, and finally to obtain a complete relationship expression. In order to obtain a better effect, the graph convolutional neural network is used in the model to predict the relationship between the target object and the surrounding objects from two directions of continuous time dimension and space dimension in the same period of time. For the relationship in the generated video clips, the structure of hypothesis tree is used to fuse them, so as to get the prediction of the visual relationship of the object in a long period of time, and the exact relationship triplet is output and expressed by using the confidence degree and other methods.

Firstly, this paper makes full use of the time and space dimension information of each video clip to construct a spatio-temporal diagram of the target entity and its relationship with each other. For the static entity relationships in video clips, the improved VRGE model based on graph neural network is used to gather the context information, so as to predict the relationship in the constructed spatio-temporal graph. For the dynamic relationship between entities, this paper predicts the relative changes of the location and appearance of entities in the space-time dimension. In addition, trajectory overlap and appearance correlation are also used to represent the degree of correlation between entities in the graph, so as to promote multi-channel information exchange.

Secondly, in order to integrate the relational triples between video clips and obtain results with higher confidence, the idea of a multi-hypothesis association MHF is used to construct multi-hypothesis tree to retain all possible correspondence relations, so as to make use of more information in the following parts. Each node in the tree represents a piece of the relationship observed in a short video clip. When the MHF processes video clips sequentially, the observed relationship fragments are selectively added to the corresponding tree as leaf nodes to update or create the hypothesis. Each path from the root node of the tree to the leaf node represents an assumption about a complete relationship instance. The greedy association algorithm used in the existing models can greedily connect the same relational triplet with high trajectory overlap in adjacent video clips, but it has some shortcomings, such as inaccurate prediction and omissions, and deviation and noise in learning and modeling long mantisseria data distribution. The multi-hypothesis correlation solves these shortcomings of the existing visual relationship prediction models on short video clips.

The main work and innovation of this paper are as follows:

1. We introduce a novel two-stage framework consists of video visual relation detection and multiple hypothesis fusion to combine the object relations and contextual information on two standard tasks: relation detection and relation tagging.
2. We proposed the deep learning model based on the graph neural network to capture the contextual information in the spatio-temporal dimension and predict short-term entity relations from video segmentation.
3. We proposed a novel algorithm based on multiple hypothesis trees to aggregate the extracted short-term relationships between entities.
4. The framework was experimented on the fully annotated ImageNet-VIDVRD, replicated, and compared with the existing methods to evaluate the performance of the model in the task of object coordinate positioning and relation detection, and the model achieved better results.

II. RELATED WORK
A. OBJECT DETECTION

According to the generation of candidate frames in the detection process, object detection can be divided into two categories: single-stage and two-stage algorithms. R-CNN series is a classic two-stage algorithm, which was first proposed by Girshick et al. [4] in 2015. Because the training process of R-CNN is complicated and the testing speed is not fast, Ren et al. [5] put forward Faster R-CNN on this basis. YOLO series is a classic single-stage algorithm. The advantage of this method lies in its fast detection speed, but its accuracy will be greatly affected by the background and surrounding objects, Redmon and Farhadi [6] proposed an improved version of YOLOv2 in 2017. In addition, Redmon also proposed a new training method, which used the joint training algorithm to act on ImageNet and COCO data sets at the same time, and trained YOLO9000, which can detect over 9000 objects in real-time. Redmon and Farhadi [7] then proposed YOLOv3 algorithm with low background false detection rate.

B. MULTI-OBJECT TRACKING

The object tracking problem can be regarded as a process of judging whether the common features of the target objects in adjacent frames match, and then constructing the location change path. In the process of time development, the target tracking technology has gradually changed from the traditional mean shift-based mode to the target tracking based on correlation filtering, and then to the target tracking algorithm based on deep learning and the target tracking algorithm based on twin network, and the technology has developed rapidly. Target tracking based on deep learning can be divided into two categories: one is the combination of features and filtering, and the combination of traditional methods and new.
methods, such as ECO [8]; One is to construct an end-to-end deep learning neural network by using convolutional neural networks, such as TCNN [9], which construct a tree structure from a limited number of CNN. The evaluation results of accuracy, robustness, and average overlap expectation (EAO) of these two kinds of deep learning target tracking algorithms show that the second kind of methods is mostly superior to the first kind, but the speed of detection and tracking of the second kind of deep learning network is not enough to achieve the goal of real-time tracking, so a target tracking technology based on twin network is proposed. The twin network has two sub-networks with the same structure, and the weights are shared among the networks. By adding different modules to the original twin network framework, Siamese Network has evolved many models with better performance. For example, Cen and Jung [10] added a full convolution network to the original twin network, thus proposing a SiamFC network structure with faster processing speed and shorter time consumption; Wang et al. [11] added a branch structure based on SiamFC and designed the SiamMask algorithm, which can ignore the differences of samples in the same category and realize the real-time tracking of targets without discrimination.

C. VIDEO VISUAL RELATION DETECTION
Visual relationship detection tasks are divided into image visual relationship detection and video visual relationship detection. Due to the complexity of the research on VidVRD and the lack of suitable data sets, it has received a lot of attention recently. Shang contributed ImageNet-VidVRD data set, which is the first dataset for video visual relationship detection. To save the workload of relation annotation, Shang and his colleagues annotated the whole video in the test set and typical video clips in the training set. In addition, they also put forward an effective three-stage detection method, including object trajectory suggestion, relationship prediction, and greedy relationship association, which has become the most widely used pipeline in VidVRD. Lu et al. [3] trained objects and predicates respectively, and then combined them to predict the relationship. Zhang et al. [12] mapped the characteristics of objects and relationships to low-dimensional space, and transformed the process of constructing relationship triplets into the process of vector translation. Zellers et al. [13] adopts two bidirectional LSTM, one for encoding the global context across the boundary region, and another for calculating and transmitting the information of the edge under the condition of prior knowledge. Because of the flexibility of graph structure, graph neural network is also applied to video relationship detection [14]. Chen et al. [15] applied the hierarchical graph matching to generate textual embeddings and aggregate matchings from video-text levels to capture global and local features. Yang et al. [16] applied attention map neural network (aGCN) to transfer context features in the whole scene for relational reasoning. Deng et al. [17] contributed a large-scale VidOR data set for VidVRD. On this dataset, Russakovsky et al. [18] applied linguistic context features and spatio-temporal features to predict predicates, and won first prize in VRU’19 (Video Relationship Understanding 2019) contest.

III. MULTIPLE HYPOTHESIS FUSION
For relatively precise relation segment detection, videos are usually split into segments of 30 frames and with 15 frames of overlap between two adjacent segments, denoted as \( S_{i1}, S_{i2}, S_{i3}, \ldots, S_{ik} \) where \( S_{ij} \) is the time segmented. For each video segment, the detected relation segments are denoted as \( O_{i1}, O_{i2}, \ldots, O_{IN}, O_{ij} \) \((1 \leq i \leq k, 1 \leq j \leq N_i)\) represents a relation segment observed in video segment \( S_i \) and \( N_i \) is the total number of observations in \( S_i \). Figure 1 shows the framework of the proposed Multiple Hypothesis Fusion (MHF) method which builds a dynamically-growing hypothesis tree for each probable video relation.

A. TARGET TRAJECTORY ACQUISITION
Given a video, it is first divided into small segments of 30 frames, and the first half and the second half coincide with the adjacent segments respectively, and then the corresponding trajectory is generated for the objects in the video segment. In the process of target tracking, target detection method of Kang et al. [19] is referred to and OP(Object Proposal) method is used. The process of obtaining the target category and trajectory can be roughly divided into two parts. The first part is to filter the sample by using the target
detector, remove the wrong part, and score the remaining choices to get the category of the target object. The second part is to use the algorithm of target tracking to get the trajectory with high confidence.

Selective Search (SS) algorithm is generally used to obtain the target category in the first part. The target detector adopted in this experiment was trained by Faster R-CNN [5] on MS-COCO [20] and ILSVRC2016-DET [18] data sets containing 35 categories of training/test images. In the second part, the open source and free machine learning algorithm library Dlib is used to track the target trajectory. For trajectories with a high degree of overlap, NMS suppression will be used to reduce repeated calculations or judgments.

**B. VRGE MODULE**

When many Video clips are used as the input of VRGE (Video Visual Relation Graph Evolution) module, the input Video clips are only Si-1, Si, and Si+1, and the input is the track feature. In order to further improve the accuracy, the current segment Si was formed into a group with the preceding and subsequent segments, and the two groups were processed separately by affinity matrix. The size of similarity matrix A is the square of the number of entities to be observed in the two segments, denoted as (2N)².

For the branch composed of two fragments, VRGE module is used to process it. The structure in the module is shown in Figure 3.

The basic model of this module is GCN, so the formula of information transfer in the GCN network layer is used: \( X \in \mathbb{R}^{2N \times d} \) is the characteristic input of this layer, \( X' \in \mathbb{R}^{2N \times d} \) is the characteristic output of this layer, \( A \in \mathbb{R}^{2N \times 2N} \) is the similarity matrix, generally the adjacency matrix or similarity matrix of the graph, \( W \in \mathbb{R}^{d \times d} \) is the parameter that needs to be optimized in the training process, \( d \) is the number of channels, \( \sigma \) is the nonlinear activation function. In the training process of deep learning, nonlinear activation functions generally include Sigmoid function, Tanh function and ReLU function. Since Sigmoid and tanh functions will have the problem of gradient disappearing and weight updating is slow in some extreme cases, ReLU is often used as the activation function.

In order to make the entity-relationship detection of the current segment more accurate, more useful information needs to be obtained from the context, so the similarity matrix A is defined in two new forms: the vIoU matrix \( A_t \) of the trajectory and the correlation matrix \( A_c \) of the appearance. These two matrices correspond to the two branches of track GCN and appearance GCN, and normalization processing is added to finally output \( X' \).

The value of each position of vIoU matrix \( A_t \) is denoted as \( A_{ti} \), where \( i \) and \( j \) (\( 0 \leq i, j \leq n-1 \)) are the serial numbers of the entities in the segment, \( T_i \) and \( T_j \) are their corresponding trajectories. The Manhattan norm is used to normalize the vIoU matrix.

Finished the two forms of similarity matrix calculation, respectively to track GCN and appearance of GCN input, do the following transform-based GCN changes finished, to avoid the characteristic value in the process of calculation is beyond the scope of the actual value, so be normalized processing, the results here using the normalization method. The matrices of the output of the two branches are obtained, which need to be added bit by bit, and then the activation function and normalization are used to obtain the output \( X' \) of the VRGE module.

The adjacent segment feature matrix is constructed as a set of \( X \in \mathbb{R}^{2N \times d} \), through many VRGE stack modules, realize the exchange of information between remote entity nodes, such as physical entities “a” respectively and the node \( b, c \) correlation is strong, but the link between \( b \) and \( c \) is not strong, so this model can help to establish information transfer between \( b \) and \( c \), and the performance of the model with the iteration, will be more accurate and more robust.

For relatively precise relation segment detection, videos are usually split into segments of 30 frames and with 15 frames of overlap between two adjacent segments, denoted as \( V = S_1, S_2, \ldots, S_i \) where \( S_i \) are sorted by time. For each video segment, the detected relation segments are denoted as \( S_i = O_{i1}, O_{i2}, \ldots, O_{iN}, (1 \leq i \leq k, 1 \leq j \leq N_i) \) represents a relation segment observed in video segment \( S_i \) and \( N_i \) is the total number of observations in \( S_i \).

Figure 2 shows the framework of the proposed Multiple Hypothesis Fusion (MHF) method which builds a dynamically-growing hypothesis the tree for each probable video relation. As shown in the Figure 4, each node in tree represents an observed relation segment and the nodes in the same level come from the same video segment \( S \). The information in each node \( O \) includes their relation triplets, trajectory (subject and object), and prediction confidence scores (subject, predicate, and object). The path of a tree from its root to a leaf node means a possible constitution of the corresponding relation, i.e. a hypothesis. During the MHF process, video segments are operated by turns. In each iteration, the hypothesis trees built from the former video segments are updated with all new observations in the current video segment by gating, scoring, and pruning process.

1) FORECAST ENTITIES AND RELATIONSHIPS

After the processing of two adjacent groups of fragments by the VRGE module, the elements in the matrix are added according to the bits, and the feature graph \( Z \in \mathbb{R}^{N \times d} \) belonging to the current fragment is extracted. \( Z \) contains the information related to the before and after fragments, which can be used to predict the category of objects and the classification of predicates. As there are many relational triplets, it is unrealistic to make predictions in the form of tuples. Therefore, subject, object, and predicate are predicted respectively. The flow chart of the prediction process is shown in Figure 3.

2) DESIGN OF LOSS FUNCTION

Deep learning and training processes, all need Loss function or cost function to help obtain a more effective model,
through multiple iterative training, the value of the Loss function is gradually decreased, and converges to a fixed value, then it means the model of adaptive parameter adjustment to a more appropriate value, take a more robust model. In this model, the use of a single classification loss function and cross-entropy, cross-entropy loss function, CE) to calculate the object in the image classification prediction vector \( V_o \) and ground-truth (i.e. artificial mark true value) between object categories TAB \( L_2 \) to losses, and use the tabbed classification and binary cross-entropy loss function (binary cross-entropy loss function, BCE) to calculate the \( V_p \) and manual annotation relationship between vector \( V_{gt} \) losses. Therefore, the accuracy of the prediction objects and predicate can be improved. In order to balance the influence brought by the two losses, they will be multiplied by the super parameters \( W_o \) and \( W_p \), and then added to get the final loss. The calculation formula is as follows: 3-10. The loss function is essential for model training. By changing the size of hyperparameters and adjusting the weight of different parts, the model can achieve convergence and improve the accuracy and robustness of the model.

### C. MULTIPLE HYPOTHESIS FUSION ALGORITHM

After the relationship instance within the clip is obtained, the relationship of the whole video needs to be detected. However, the Multi Hypothesis Fusion (MHF) algorithm outputs the short-term relationship instance of the whole video, which fuses the relationship triples within multiple clips and outputs the final relationship instance. \( V = S_1, S_2, \ldots, S_k \), S stands for video clips of fixed length, generally taking 30 frames, sorted according to time. For each video clip, the detected relationship is expressed as \( S_i = O_{i1}, O_{i2}, \ldots, O_{iN_i}, O_{ij}(1 \leq i \leq k, 1 \leq j \leq N_i) \) notes the relationship segment observed in the video segment \( S_i \), where \( N_i \) is the total number of observations for \( S_i \). Figure 2 is a block diagram of the multi-hypothesis fusion algorithm.

Multi-hypothesis fusion algorithm will build a hypothesis tree for each video relationship segment, which depends on the new video segment and the previous hypothesis tree, that is, the hypothesis tree is dynamically growing. As shown in Figure 4, each node in the tree represents an observed relationship segment \( O_{ij} \), and the nodes at the same level are from the same video segment \( S_i \). The information in each node includes the relationship triad \(<\text{subject, relation, object}>\), the trajectory of subject and object, and the prediction confidence score of each item in the triad. All new nodes in the current video clip are updated to the hypothesis tree constructed by previous video clips by selecting the process of build, score, and pruning, and the continuous iteration is carried out until the construction of all video clips is completed. This section details the process of building the hypothesis tree and iterating over it.

1) **SUPPOSE THE CONSTRUCTION OF THE TREE**

When entering a new relationship between segments, because each fragment with relation triples, observe first new...
ILLUSTRATION OF VIDEO VRD BASED ON GRAPH NEURAL NETWORK

**Figure 4.** An overview of the multi-hypothesis fusion method. The relation hypothesis trees generated by observations in the former video segments are updated with new observations in the current video segment during each iteration. The results after each iteration will act as “former hypothesis trees” in the next iteration. $O_j$ presents the $j$-th relation segment observed in video segment $S_i$. Fake nodes are illustrated with green color. In the hypothesis trees at the top-most part, the optimal path of each tree is denoted by bold line while the paths being pruned are denoted by dotted lines.

fragments that contain the same relation triples and leaf nodes, see them as a pair, the node may have many, so you also need to calculate, the connection between each pair of similarity is higher than the threshold value of artificial nodes on, to achieve the connection. Connection similarity is used to measure the tightness of the connection between the new segment and each leaf node. The calculation method is as follows:

\[
S_{\text{con},s} = \alpha \times \text{vIoU}_s + \beta \times s_s. \tag{1}
\]

\[
S_{\text{con},o} = \alpha \times \text{vIoU}_o + \beta \times s_o. \tag{2}
\]

Since the relational triples of node pairs are the same, the subject and object in the triplet are respectively calculated, namely, $S_{\text{con},s}$ and $S_{\text{con},o}$, which represent the connection similarity. IoU is the intersection and union ratio, while vIoU refers to the total IoU of the two different tracks of the subject and vIoU$_o$ corresponds to the total IoU of the object. $S_s$ and $S_o$ represent the host-guest confidence of nodes corresponding to the new segment, and $\alpha$ and $\beta$ are super parameters.

The calculation of connection similarity takes advantage of geometric information and uses confidence to increase the reliability of the prediction, so the connection score is more robust. Because predicate prediction accuracy in relational triples is relatively low, the confidence score of a predicate is not included in the calculation of connection similarity force. In the process of connecting nodes and preliminary construction of the hypothesis tree, those that remain isolated will become the root of the new tree, as shown in figure O44.

On the other hand, when an existing hypothetical leaf node in the tree does not receive any connection in the iteration, which means that loss detection may have occurred, a virtual node containing the same information as this leaf node is connected to it as its child node and becomes the new leaf node, as shown in the Figure 4.

2) CALCULATION OF RELIABILITY

When the initial wiring of the hypothesis tree is completed, each relational hypothesis tree may have multiple paths that represent multiple hypotheses that are updated to the corresponding video relationship of the current video clip. A path score is designed to measure the reliability of each hypothesis, and the less reliable hypotheses are removed to facilitate subsequent operations.

\[
S_{\text{rel}} = \frac{S_{\text{con},s} + S_{\text{con},o} + \gamma \times S_p}{2 + \gamma}. \tag{3}
\]

\[
S_{\text{path}} = \text{AVG}(S_{\text{rel}}). \tag{4}
\]

Equation 3 represents the score of each node on the path, which is the weighted average of the connecting similarity score of the subject and object and the predicate prediction score and contains the information of the previous node. $S_p$ represents the confidence score of the predicate. Because the root node has no previous node used to calculate the join score, its score is set to its triple confidence score, equal to $S_s \times S_p \times S_o/10^f$, where $f$ is a scaling factor used to make
the triple confidence score of the same order of magnitude as the node score calculated by Equations 1.

Formula 4 represents the score of the entire path, which is the average score of all nodes on the path. This approach takes into account the characteristics of long-term connection and the detection confidence and models the reliability of forming the corresponding hypothesis video relationship. It comprehensively measures the reliability of the formation of assumptions about video relationships.

3) PATH SELECTION AND ITERATION
The existing hypothesis tree will be pruned after the score of each path is obtained. The process of constructing a hypothesis tree is the process of gathering effective information in the new video clips. When there is enough information to make a judgment, some hypotheses with low reliability may be proved to be wrong, so they will be deleted. Since the construction of the hypothesis tree does not take into account the second half of the video, it is impossible to definitively prove that the hypothesis is necessarily wrong, so the target of pruning is limited to the branches that occur before N video clips.

Since there are multiple hypothesis trees, conflicts may occur in the selection process of the optimal path, so the selection of the optimal path in the global scope is called the formation of a global hypothesis. The formation of the global hypothesis can be expressed as the following optimization problem:

\[ \max z \sum_{j=0}^{N_1} z_j + \sum_{j=2}^{N_2} z_j + \cdots + \sum_{j=0}^{N_K} z_j \cdot (5) \]

\( s \) represents the score of the corresponding path, and \( z \) is the binary variable. When \( j = 0 \), it means that the path does not contain the observation in the video clip \( S_i \). In VidVRD, after the generation of the hypothetical tree, the number of trees may be more than the actual number of relationships, which means that only a part of the tree can produce the correct video relationship we need. To select paths in these trees, we should be greedy to choose high-resolution paths that reflect high reliability, so the greedy algorithm will be utilized in the selection of the global optimal hypothesis. With the result of global hypothesis formation, the pruning step can be accomplished easily. Here we set \( N \) as 2, and ABCD in the figure is the calculated value. Assuming that the calculated value of \( C \) is the minimum, then the non-optimal branches formed before the two video clips shown in the figure will be pruned in each tree. After this two-scan pruning process, only the last level in each tree branches, and each previous level has only one node.

4) THE FORMATION OF VIDEO RELATIONSHIPS
After the multi-hypothesis fusion algorithm has processed all the video clips, the final hypothesis tree will be used to generate the final relational results. Knowing that the global hypothesis forms the optimal hypothesis for the selection, we simply need to concatenate the nodes in the selected path and transform them into a single video relationship.

When the result is generated, virtual nodes in the path are skipped. For two adjacent nodes in a path that skips a virtual node, if their trajectories overlap, we can connect them by averaging their bounding boxes throughout the overlap. Otherwise, if their trajectories do not overlap, which means that there are virtual nodes and miss detection between them, we use linear interpolation to manually generate the miss trajectories. We used the path score in Equations 4 for the evaluation.

IV. EXPERIMENT
A. DATASET
The dataset used in our experiment is ImageNet-VidVRD. The ImageNet-VIDVRD dataset is based on the contents of the ILSVRC2015-VID [18] dataset, from which 1000 videos are collected, and the subject and object categories and corresponding tracks in these videos are marked. Visual relationships are labeled under 35 object categories and 132 predicate categories. In the training process, 80% of the whole dataset was used for training and 20% for testing, that is, 800 training videos and 200 test videos. To save the workload of relational labeling, only typical clips of the video were labeled in the training set, while the whole video was labeled in the test set. The statistical data of the dataset are shown in Table 1.

| TABLE 1. ImageNet-VidVRD. |
|---------------------------|
|                        | training set | test set |
| video                    | 800          | 200      |
| subject/object category  | 35           | 35       |
| predicate category       | 132          | 132      |
| relation triplet         | 2961         | 1011     |
| visual relation instance(video) | -      | 4835     |
| segment                  | 15146        | 3202     |
| labeled segment          | 3033         | 2801     |
| visual relation instance(segment) | 25917 | 29714    |

B. EVALUATION INDICATORS
On both datasets, we evaluate our method on two standard tasks as used in [21] relation detection and relation tagging. The input of the relationship detection task is the video clip, and the relationship triplet detected in the clip and the corresponding subject/object trajectory in the video is obtained. If the detected relational triplet is the same as the true labeled relational triplet and the oblivious of both subject and object trajectories is above the threshold, the relationship is considered to be accurate. The vIoU threshold was set at 0.5 and the quantitative assessment was performed using the mean accuracy (mAP) and recall rate Recall@K (K = 50, 100). The relational marking task outputs all the relational triples of the entire video, eliminating the requirement of object positioning and only paying attention to the precision of relational triples. If the detected relational triplet is true, the relationship can be considered correct. For this task, accuracy Precision@K was used as the evaluation metric, and
since the average number of relational triples in the dataset in each video clip was 10.34, K was set to = 1,5,10.

C. ABLATION STUDY

This experiment design is the overall framework of reference for the Shang et al. [21] three steps of the model is put forward, so the results and the experiment of Shang performance comparison, and compared the proposed by Qian et al. [22] relationship prediction algorithms based on neural network design, and compare the Shang and Qian in multiple hypothesis fusion algorithm (M HF) performance before and after the change.

In addition, for the VRGE, to verify that both GCN branches are helpful to the performance of relationship detection, the following results are obtained using a controlled experiment, as shown in Table 2.

| TABLE 2. Two parts of VRGE module. |
|----------------------------------|
| Relationship detection | Relationship labeling |
| mAP | R@50 | R@100 | P@1 | P@5 | P@10 |
| appearance GCN | 6.08 | 7.10 | 12.48 | 52.00 | 37.50 | 25.85 |
| track GCN | 7.45 | 9.17 | 15.31 | 55.00 | 39.05 | 27.00 |
| appearance+track GCN | 8.04 | 9.23 | 16.06 | 57.50 | 40.50 | 28.00 |

The trajectory vIoU matrix represents how two objects are related geometrically, and the appearance correlation matrix represents the intrinsic correlation of two objects. It can be observed that the locus GCN branch using the vIoU matrix performs better than the appearance GCN branch using the appearance correlation matrix because the spatial geometric information is more important in describing the relationship between relative motion and position. However, the combined effect was 0.75% better on the detection task (mAP) than the track GCN alone. Therefore, it is verified that both branches are helpful to the performance of relationship detection.

In the construction process of the detection model, multiple VRGE modules can be used continuously, and the experimental effect is shown in Table 3. The results show that when the number is less than 4, the structure of the model can benefit from the increase of complexity, and the overall performance of the model is the best on the three-layer detection and labeling tasks. When the number of layers is increased to 4 layers, the overfitting problem of performance degradation will be encountered. So in the other comparison experiments, the default VRGE was set to three.

| TABLE 3. The numbers of VRGE. |
|-------------------------------|
| The Number of VRGE | Relationship detection | Relationship labeling |
| mAP | R@50 | R@100 | P@1 | P@5 | P@10 |
| 0 | 4.27 | 5.16 | 8.28 | 34.50 | 20.66 | 15.38 |
| 1 | 6.13 | 7.35 | 13.02 | 46.50 | 31.85 | 22.10 |
| 3 | 8.03 | 9.13 | 16.26 | 57.50 | 41.00 | 28.20 |
| 4 | 6.87 | 8.07 | 14.67 | 58.85 | 39.30 | 28.15 |

In the multi-hypothesis fusion algorithm, there are some super parameters. Through the experimental test, the final setting is $\alpha = 0.6$, $\beta = 0.4$, and $\gamma = 0.6$, and $N = 2$. During the construction of the hypothesis tree, the connection similarity needs to be higher than the artificial threshold, which is set to 0.5. In the process of path selection, to reduce computational complexity, save storage space and facilitate subsequent training, a simple pruning operation will be performed immediately after each iteration, so that there are at most 5 leaf nodes in each hypothetical tree. At the same time, during the experiment, disconnected connections were mistakenly connected, so self-checking operations were added after each iteration to exchange calculation time for higher accuracy. For a hypothetical tree with no new non-virtual leaf node connections for a long time, we determine that it has completed the maximum end-to-end extension, so we generate an instance of the video visual relationship corresponding to the optimal path of the tree. This “long time” demarcation is achieved by setting 3 to the maximum number of contiguous virtual nodes.

In the experiment, we use the object trajectory of Shang as the input, which is also used in the video visual relationship detection experiment of Qian et al. The two research teams used the same relational feature determination method, that is, the category of two objects in the relational triad and their trajectory characteristics were used to determine the predicate representing the relationship. There are two kinds of object-related features: one is classification feature, and the other is Dense Trajectory features obtained by IDT (Improved Dense Trajectory) algorithm. The relation between objects, namely the relation characteristic, is related to the relative position, size, and motion process between the subject and object. Shang used three predictors for subject-object and predicate and trained relational triples by combining paired loci with manual labels that overlap vIoU over 0.5. The top 20 predicted results for each track pair and the top 200 results for each video clip were retained for post-scoring selection. Since the VIDVRD confidence score is one order of magnitude higher, the scaling factor f is set to 1 in all the reliability calculation stages. Qian extracts the relational triples and target features of three adjacent video clips in the experiment, which are used as the input of the ST-GCN module for relational prediction, and sets the number of trajectories reserved after each video clip prediction as to the first five. For the prediction of the target category and the prediction of the predicate category, different linear transformation layers are adopted to make full use of the context and the surrounding object information with a higher correlation degree to achieve a more accurate prediction.

Based on the experiments of Shang and Qian, we improve the graph convolutional neural network algorithm in the relationship prediction stage and use the multi-hypothesis fusion algorithm to fuse the visual relations corresponding to short clips into the dynamic visual relations of the whole video. Finally, the performance index was obtained by comparing the two tasks of relationship detection and relationship marking.

D. RESULTS OF EXPERIMENT

In the process of the experiment, we have adopted two groups of control experiments, including the comparison between
Shang’s and Qian’s respectively before and after the combination of VRGE-MGF algorithm, and the comparison between Qian’s+ Siamese and Qian’s+ VRGE-MGF, to get the scores under the two evaluation criteria of relationship detection and relationship marker, as shown in Table 4 and 5.

| Control experiment | Relationship detection | Relationship labeling |
|--------------------|------------------------|-----------------------|
|                    | mAP       | R@50      | R@100     | P@5      | P@10     | P@50     | P@100    |
| Shang’s            | 8.43      | 5.32      | 6.17      | 43.00    | 28.10    | 20.25    |
| Qian’s             | 13.78     | 7.13      | 8.45      | 59.50    | 40.20    | 27.55    |
| Shang’s+VRGE-MGF   | 15.65     | 7.34      | 8.42      | 60.00    | 26.35    | 17.45    |
| Qian’s+VRGE-MGF    | 18.89     | 9.33      | 10.18     | 57.20    | 41.10    | 29.05    |

For the first group of control experiments, Shang’s and Qian’s were compared before and after the combination of the VRGE-MHF algorithm. These all use the same target trajectory, the difference lies in the relationship prediction and fusion stage. It can be seen from the data that MAP improved a lot in the evaluation results related to relationship detection, especially for Shang’s experiment. The recall rate R@50 and R@100 both had a performance improvement of about 2%. However, in terms of accuracy, Shang’s combined with VRGE-MHF algorithm is not excellent enough. For Qian’s work, the performance of P@5 and P@10 is slightly better, indicating that VRGE-MHF has a better detection effect on video clips with a large number of relational triples.

Comparison of the second control experiment: Qian’s+ Siamese and Qian’s+ VRGE-MHF. Siamese uses twin networks to achieve the fusion and association of relationship fragments, while VRGE-MHF uses a dynamically updated hypothesis tree. As can be seen from the scores of experimental indicators, VRGE-MHF has a better effect to detection and relation labeling, and both recall rate and accuracy are significantly improved. In terms of resource consumption, the use of VRGE-MHF will definitely increase memory and computing resources. Compared with the commonly used greedy relational association in existing work, VRGE-MHF consumes a small number of resources, but it can obtain better experimental results. As can be seen from the above comparative experiment, under the task of relationship detection and evaluation, the VRGE-MHF method can combine different prediction models, and the effect is significantly improved. Under the relation mark assessment tasks, VRGE-MHF correlation method because of it is designed based on dynamic growth tree, all can real-time adjust the confidence score, to improve the accuracy, but because of the existence of pruning operation may appear correct relation triples excluded, deterioration in part of the evaluation results.

In the first set of examples, VRGE-MHF generated more correct relationships and achieved better confidence scores. In the second set of examples, multiple objects in similar states are present in the image, and they have similar geometric positions, which may lead to confusion during association, resulting in no correct relational triples. However, the idea of multiple assumptions will retain the possibility of high confidence in the process of prediction, and make decisions after acquiring more information in the future, so the accuracy rate is higher and the effect is better. Experiment process, also the relationship between forecasting is not smooth, such as when the subject or predicate the category of the judge when something goes wrong, because the build suppose that process is not included in the modified node of the tree, so the final output of triples the relationship must be a mistake, that is to say for category input error, without the ability to correct. In addition, due to the variability of predicates, complex predicates contain a variety of meanings. For example, “run towards” includes the direction of action and behavior, which makes the prediction of predicates more difficult.

V. CONCLUSION

In our study, we proposed a model based on the graph neural network to predict short-term relationships, and we also released a novel relation association method MHF for VidVRD. VRGE generates lots of short-term visual relationships, and the output is the input of MHF. It generates dynamic relation hypothesis trees to track and maintain multiple hypotheses of relations. A variety of comparative tests on the ImageNet-VidVRD dataset indicates the effectiveness of our method and proves that the idea of multiple hypotheses can indeed play an important role in VidVRD. However, our method still has much room for improvement on both relation detection and tagging tasks. And the task of VidVRD can be improved with several proposals as following: (1) build a data set with higher resolution and richer object categories and relationships, and make more comprehensive annotation; (2) A more comprehensive evaluation method is adopted to expand the scenarios in which the model can be applied, so as to make achievements in both academic and industrial circles. Through the above several improved perspectives, the object visual relationship detection research is more effective, more valuable. Through the above several improved perspectives, the object visual relationship detection research will be more effective, more valuable.

REFERENCES

[1] T. Bouwmans, “Traditional and recent approaches in background modeling for foreground detection: An overview,” Comput. Sci. Rev., vol. 11–12, pp. 31–66, May 2014.
[2] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” in Proc. 6th Int. Conf. Learn. Represent. (ICLR), Vancouver, BC, Canada, Apr./May 2018, pp. 1–12.
[3] C. Lu, R. Krishna, M. S. Bernstein, and L. Fei-Fei, “Visual relationship detection with language priors,” in Computer Vision—ECCV 2016 (Lecture Notes in Computer Science), vol. 9905, B. Leibe, J. Matas, N. Sebe, and M. Wellinger, Eds. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 852–869.
[4] R. Girshick, F. Iandola, T. Darrell, and J. Malik, “Deformable part models are convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA: IEEE Computer Society, Jun. 2015, pp. 437–446.

[5] S. Ren, K. He, R. B. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Ann. Conf. Neural Inf. Process. Syst., C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., Montreal, QC, Canada, Dec. 2015, pp. 91–99.

[6] J. Redmon and A. Farhadi, “YOLO9000: Better, faster, stronger,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA: IEEE Computer Society, Jul. 2017, pp. 650–655.

[7] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” CoRR, vol. abs/1804.02767, pp. 1–6, Apr. 2018.

[8] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, “ECO: Efficient convolution operators for tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA: IEEE Computer Society, Jul. 2017, pp. 6931–6939.

[9] H. Nam, M. Baek, and B. Han, “Modeling and propagating CNNs in a tree structure for visual tracking,” CoRR, vol. abs/1608.07242, pp. 1–10, Aug. 2016.

[10] M. Cen and C. Jung, “Fully convolutional Siamese fusion networks for object tracking,” in Proc. 25th IEEE Int. Conf. Image Process. (ICIP), Athens, Greece, Oct. 2018, pp. 3718–3722.

[11] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, “Fast online object tracking and segmentation: A unifying approach,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA: Computer Vision Foundation, Jun. 2019, pp. 1328–1338.

[12] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, “Visual translation embedding network for visual relation detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA: IEEE Computer Society, Jul. 2017, pp. 3107–3115.

[13] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, “Neural motifs: Scene graph parsing with global context,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Salt Lake City, UT, USA: IEEE Computer Society, Jun. 2018, pp. 5831–5840.

[14] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” IEEE Trans. Neural Netw., vol. 20, no. 1, pp. 61–80, Jan. 2009.

[15] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, “Fine-grained video-text retrieval with hierarchical graph reasoning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 10635–10644, doi: 10.1109/CVPR42600.2020.01065.

[16] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, “Graph R-CNN for scene graph generation,” in Proc. 15th Eur. Conf. Comput. Vis., in Lecture Notes in Computer Science, vol. 11205, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Munich, Germany: Springer, Aug. 2018, pp. 690–706.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Miami, FL, USA: IEEE Computer Society, Jun. 2009, pp. 248–255.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015.

[19] K. Kang, W. Ouyang, H. Li, and X. Wang, “Object detection from video tubeslet with convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA: IEEE Computer Society, Jun. 2016, pp. 817–825.

[20] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Computer Vision—ECCV 2014 (Lecture Notes in Computer Science), vol. 8693, D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Zürich, Switzerland: Springer, Sep. 2014, pp. 740–755.

[21] X. Shang, T. Ren, J. Guo, H. Zhang, and T.-S. Chua, “Video visual relation detection,” in Proc. 25th ACM Int. Conf. Multimedia (MM), Q. Liu, R. Lienhart, H. Wang, S. K. Chen, S. Boll, Y. P. Chen, G. Friedland, J. Li, and S. Yan, Eds., Mountain View, CA, USA, Oct. 2017, pp. 1300–1308.

[22] X. Qian, Y. Zhuang, Y. Li, S. Xiao, S. Pu, and J. Xiao, “Video relation detection with spatio-temporal graph,” in Proc. 27th ACM Int. Conf. Multimedia (MM), L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, and W. T. Ooi, Eds., Nice, France, Oct. 2019, pp. 84–93.