ABSTRACT

Video visual relation detection, which aims to detect the visual relations between objects in the form of relation triplet <subject, predicate, object> (e.g., “person-ride-bike”, “dog-toward-car”, etc.), is a significant and fundamental task in computer vision. However, most of the existing works about visual relation instances are focused on static images. Modeling the non-static relationships in videos has drawn little attention due to lacking large-scale video dataset support. In our work, we propose a video dataset named Video Predicate Detection and Reasoning (VidPDR) for dynamic video visual relation detection, which consists of 1,000 videos with dense manually dynamic labeled annotations on 21 object classes and 37 predicates classes. Moreover, we propose a novel spatio-temporal feature extraction framework with 3D Convolutional Neural Networks (ST3DCNN), which includes three modules 1) object trajectory, 2) short-term relation prediction, and 3) greedy relational association. We conducted appropriate experiments on public datasets and our own dataset (VidPDR). Results demonstrate that our proposed method has a great improvement in comparison to the state-of-the-art baselines.

INDEX TERMS

Computer vision, 3D convolutional neural network, video visual relation detection.

I. INTRODUCTION

As a bridge between dynamic object detection and textual information in video information retrieval, traditional video visual relation detection aims to explore the non-static interaction knowledge among co-occurrent objects in the video. A visual relation instance can be described as triplet <subject, predicate, object> which contains comprehensive semantic understanding by providing a pair of objects and the specific interaction. These detected visual relation instances can be widely utilized in many video applications, such as video response systems [1], [2], video caption [3]–[5], video summarization [6]–[8], video retrieval [9]–[11], etc.

Compared with visual relation detection in the still image field, recognizing visual relation in videos is a more complicated and challenging research direction. The first problem is that this research lacks support for large-scale datasets. It is because constructing a large-scale dataset for VidVRD requires lots of manpower and material. The second problem is that traditional image visual detection methods [12], [13] are not suitable for the video visual system. The video is composed of many frames which are considered still images. The existence of time dimension leads the localization of the same object called spatial-temporal localization can change over time. What’s more difficult is, due to the movement of multiple objects in the video will be superimposed, visual relations between objects could be more complicated and variable. In addition, the temporal-based object relations will be applied in video visual relation detection and increase the difficulty of the research. To tackle the second problem, several existing well-designed methods [14]–[16] divided the task flow into three stages, 1) object trajectory, 2) short-term relation prediction, and 3) greedy relational association. We used a baseline named VidVRD [14] as an example to introduce the general flow of our study. For the first stage, the image segmentation module uses the fixed-length window to segment the video into many segments of the same length. Second, a model based on object trajectory and relation prediction is applied to perform video relation detection in segments. Finally, aggregating these segments greedily generates the results of the video relation detection. After that, Tsai et al. proposed GSTEG [16] to replace the relation prediction method with the above-mentioned baseline. This method utilized a fully connected graph structure and an
energy function with trainable hyperparameters to model the spatial-temporal of object relations in the given video and achieved great performance. Qian et al. introduced the graph convolutional network [15] and utilized the spatial-temporal context to obtain better effects.

In our study, we propose a video dataset named Video Predicate Detection and Reasoning (VidPDR) for video visual relation detection, which consists of 1,000 videos with dense manually dynamic labeled annotations on 21 object classes and 37 predicates classes. Compared with previous datasets, our dataset has richer predicate categories and adds comparison predicate annotations such as faster-than, smaller-than, etc., which increases the difficulty for prediction. In addition, to improve the efficiency of the methods, we propose a novel framework with 3D Convolutional Neural Network named ST3DCNN to extract additional discriminative spatio-temporal features in the given video. The novelty of our method is divided into two parts. First, the Spatio-temporal features were generated by a 3D convolutional kernel instead of manual feature extraction to improve the prediction performance of short-term relations between pairs of objects. The 3D CNN feature extraction network will extract more abundant features for relation detection. Second, we associated all the short segments into complete relation instances. To verify the reasonability and effectiveness of our proposed feature extraction method, we conducted experiments on the new dataset (VidPDR) and two public datasets, ImageNet-VidVRD [14] and VidOR [17]. The main contributions of our work are summarized as follows:

- We contribute a video dataset named VidPDR which contains 1,000 videos with dense manually dynamic labeled annotations, providing the data resource for modelling the non-static visual relationships.
- We propose a dynamic feature extracting model with a 3D Convolutional Neural Network named ST3DCNN to extract distinguishing features to detect visual relations more effectively.
- We demonstrated that our proposed model achieved better performance in comparison to VidVRD in video visual relation detection.

II. RELATED WORK

A. VIDEO VISUAL RELATION DETECTION

Image visual relation detection consists of detecting objects and interactions between different objects. Ren et al. proposed Faster R-CNN [18] to identify the location of multiple objects in static images. After that, researchers pay attention to mining latent knowledge which is hidden in high-order relationships among the detected objects. Li et al. [19] proposed a model with a guided message passing structure to explore visual interactions and construct an information flow for object detection. Zhang et al. introduced a relationship proposal network [20] to address the limitation of the inefficient performance for detecting all objects in the scene. Khan et al. [21] introduced a series of multi-scale detectors for geo-spatial object detection in high-resolution satellite images.

Compared with the task mentioned above, video visual relation detection, as a novel video visual understanding task, is attracting a lot of concerns. The video captioning tasks [3]–[5] and video dialogue system [1], [2] which provide more advanced scenarios can use the visual relation instances to achieve good benefits of video object detection. Shang et al. proposed a general pipeline framework that encodes features by deep learning model and conducted the experiments on the first large-scale dataset (VidVRD) [14]. Many researchers also work on different research directions. In order to extract high availability discriminative features, Sun et al. [22] improved the accuracy of trajectories of objects with multiple methods including Flow-Guided Feature Aggregation(FGFA) [23], Seq-NMS [24], and KCF tracker [25]. To further encode the visual relation interactions in videos, Tsai et al. [16] introduced a graph algorithm with a gate structure whose vertices and edges indicate objects and the relationships among them, respectively. With the rise of graph neural networks, Qian et al. [15] followed the tendency and applied graph neural networks with one convolutional layer to learn the Spatio-temporal features. Nonetheless, most works mentioned above consist of multi-steps aim to implement the object trajectories and intermediate representations. Such a multi-step operation may lead to a waste of computing resources and memory. On the contrary, the feature extraction model built by 3D Convolutional Neural Network fully fits the experimental scene of the video mathematically, and it can simplify the data processing and extract efficient video representations.

B. 3D CONVOLUTIONAL NEURAL NETWORK

In the past decades, most of the researchers applied models based on 2D Convolutional Neural Networks to obtain the image features. These features which are encoded through the last fully connected layer of the network can achieve satisfactory performance in image detection. However, such image-based representations which lack the temporal information between adjacent frames in videos are not suitable for the video-based problem. To tackle this problem, Ji et al. first attempted to migrate the theory of 3D Neural Network [26] to video object relation detect task. Tran et al. designed a widely used 3D Convolutional Neural Network (C3D) [27] for extracting features from video segmentation in large-scale video datasets. As one of the earlier 3D Convolutional Neural Networks, C3D proposed a special size convolutional kernel for convolutional layers and leveraged 3D max-pooling to aggregate the features with the dimension of time. However, the complicated structure leads the C3D to fit massive hyperparameters and consume a lot of computing resources. Therefore, Pseudo-3D Residual Network (P3D) [28] simulates 3D convolutional kernel with 2D spatial convolutional kernel plus 1D temporal convolutional kernel and integrates the design into a deep learning framework. The main idea of Pseudo 3D CNN is
M. Qu et al.: Video Visual Relation Detection via 3D Convolutional Neural Network

FIGURE 1. The overview of video frames and overlapped segments.

FIGURE 2. The framework of our method.

to decouple the 3D convolutional neural network which has a complex structure into a 2D spatial convolutional layer and a 1D temporal convolutional layer and utilize the object relationships from images.

Most of the existing 3D Convolutional Neural Networks [27]–[29] have been applied in action recognition and video classification tasks etc. However, many works still convert the video into images in video visual relation detection and give attention to the 2D convolutional neural networks method. Hence, it is an inevitable trend to transform a 2D model into a 3D model by adding a time dimension.

III. METHOD

We summarize Shang et al. [14] proposed methods to detect visual relation instances in videos and introduce a greed association algorithm to generate the comprehensive visual relation. Video analysis process consumes a lot of computing resources and memory overhead. Existing works tend to study latent knowledge from short videos and migrate the strategy to obtain high-order comprehensive information from long videos. The theoretical support is that the simple visual relationship can be extracted quickly and effectively in short videos, and the complex visual relationship can always be inferred from the simple relationship. On this basis, we propose the object trajectory detection module and a relation predict module. In the stage of object trajectory detection, we detect the object from every frame in the video with the FGFA model [23] based on ResNet-101 [30] as the benchmark. Then we use the greedy relational association algorithm [14] to aggregate the results of object detection from all video frames and generate trajectories for each object. In the stage of relation prediction, we extract some more abundant and comprehensive Spatio-temporal features by a 3D Convolutional Neural Network to improve the performance of relation prediction. The overview of our proposed framework is shown in Figure 3,2. As we can see in Figure 2, the 3DCNN feature extraction network is behind the object trajectories stage to extract spatial, visual and temporal features for relation detection.

A. OBJECT TRAJECTORY DETECTION

Object trajectory detection is an important sub-task in computer vision, which determines the performance of a computer’s ability to understand behavior. An accurate object trajectory detection algorithm can commendably improve the performance of the video visual relation detection method, while a poor object trajectory detection algorithm will have a terrible effect on the performance of the video visual relation detection method. Hence, to prevent the number of generated relation candidates in the combination from being too large, we need a limited number of but high-quality object trajectories. Then, we fuse several available advanced methods about video object detection and aim to obtain more accurate object trajectory detection instances.

At first, the FGFA model [23] based on ResNet-101 [30] was trained as the pre-trained model with ImageNet [31]. Then, we use the trained FGFA model to detect individual objects. During the object detection process, we filtered out the result of object detection whose confidences are less...
than 0.01 in our experiments. On this basis, we introduced a strategy that consists of two steps (tracking and detection). To obtain the short preliminary trajectories after filtering results of object detection, we employ the Seq-NMS which consumes little computing resources. Since the associated standard is only based on the overlap in Seq-NMS, it is difficult to generate the high-quality object trajectory. For example, If an object in a video moves quickly, then the IoU of the bounding box we set cannot capture the exact position of the object on time, and our tracking effect will become very poor, or even the object can’t be tracked directly. Therefore, we introduce new connecting mechanism that minimizes the error rate of the bounding box. Two bounding boxes \( B_i \) and \( B_{i+1} \) categorized as \( C_i \) and \( C_{i+1} \) on the \( i \) and \( i+1 \) frame are connected, association rules are as follows:

- The bounding boxes belong to the same categories.

\[
C_i = C_{i+1} \quad (1)
\]

- The overlap between the two bounding boxes is bigger than the threshold.

\[
\text{IoU}(B_i, B_{i+1}) > \alpha \quad (2)
\]

- The difference between the size of the two bounding boxes is less than a certain threshold.

\[
\left| \frac{B_i^h}{B_i^w} - \frac{B_{i+1}^h}{B_{i+1}^w} \right| < \beta \quad (3)
\]

where \( B_i^h \) and \( B_i^w \) denotes the height and width of the bounding boxes. We set the \( \alpha \) and \( \beta \) to 0.8 and 0.3, respectively. \(|\cdot|\) is an absolute value operation.

In Seq-NMS, since the length of the frame in the video is limited, the short-term object motion trajectory we obtain is not comprehensive. We track the position of the head and tail of the identified object and then complete the connection of the object. We use CPU parallel to complete this operation to reduce time consumption. According to the predetermined threshold (0.05 in our experiments), we screen out object trajectories that meet the requirements. In our experiments, 20 object trajectories instances that achieved the best performance are saved for the following relation prediction.

| Method               | R@50 | R@100 | mAP  | P@1   | P@5   | P@10  |
|----------------------|------|-------|------|-------|-------|-------|
| ViDiVRD [14]         | 5.54 | 6.37  | 8.58 | 42.58 | 28.54 | 20.35 |
| MFF [22]             | 7.02 | 8.98  | 8.93 | 43.75 | 30.18 | 21.73 |
| 3DRN [32]            | 5.53 | 6.39  | 14.68| 45.28 | 36.69 | 22.42 |
| GSTEG [16]           | 7.05 | 7.67  | 9.52 | 50.66 | 38.13 | 27.83 |
| ViDiVRD+MHA [33]     | 7.40 | 8.58  | 15.71| 47.13 | 33.77 | 25.95 |
| VRD-GCN [15]         | 7.43 | 8.75  | 14.23| 57.42 | 40.92 | 28.07 |
| ST3DCNN              | 7.61 | 8.96  | 9.04 | 56.87 | 41.05 | 28.97 |

B. RELATION PREDICTION

We divide the video into frame segments of the same length and overlap segments with the same number of frames between every two frames. In this experiment, we set the length of the frame segment to 32, and the overlap length to 16, so that half of our frame segments overlap. And for each video frame, we resize it to 160 × 160. Then the input dimension of the 3DCNN feature extraction network is 3 × 16 × 160 × 160, in which 3 represents that each frame has 3 channels. In our experiments, P3D is utilized as the benchmark of the 3DCNN feature extraction network. Our model uses the operation of 3D convolution to capture the time and space information in the video. The illustration of the 3DCNN feature extraction network is shown in Figure 3. By concatenating the features extracted from the network mentioned above and the features provided by [14] and [22], the input of the short relation detection module is more comprehensive.

Using the concatenated features as input, we use a fully connected network to predict the short-term relation.

IV. EXPERIMENT

In the stage of video visual relation, the input and output of our proposed model are a complete video, and the video visual relations are based on detected objects. For each output result, when each obtained visual relationship triplet is the same as the object trajectory of the object annotated in the video, and the bounding box of the object in the result coincides with the ground truth very high, we consider this result is correct. Following the set of [14], vIoU indicates the voluminal intersection over the union of two object trajectories and we set the overlapping threshold of vIoU to 0.5 in our experiments.
TABLE 2. Performance comparison of different methods on VidOR dataset for video visual relation detection.

| Method     | R@50  | R@100 | mAP  | P@1   | P@5   | P@10  |
|------------|-------|-------|------|-------|-------|-------|
| VidVRD [14]| 5.24  | 6.14  | 8.17 | 40.55 | 25.45 | 17.64 |
| RELAbuilder [34]| 1.58 | 1.85 | 1.47 | 32.06 | 14.50 | 15.55 |
| MAGUS Gamma [22]| 6.89 | 8.83 | 6.56 | 43.17 | 27.11 | 21.81 |
| VRD-STGC [35]| 8.21 | 9.90 | 8.85 | 44.62 | 29.20 | 23.92 |
| MFF [22]| 6.89 | 8.83 | 8.56 | 42.09 | 26.11 | 20.17 |
| 3DRN [32]| 2.58 | 2.75 | 2.47 | 31.90 | 13.05 | 16.59 |
| ST3DCNN| 6.98 | 8.86 | 8.84 | 45.88 | 31.33 | 24.43 |

TABLE 3. Performance comparison of different methods on VidPDR dataset for video visual relation detection.

| Method     | R@50  | R@100 | mAP  | P@1   | P@5   | P@10  |
|------------|-------|-------|------|-------|-------|-------|
| VidVRD [14]| 6.04  | 7.21  | 9.13 | 36.02 | 20.81 | 17.78 |
| MFF [22]| 7.68 | 9.53 | 9.56 | 39.72 | 22.67 | 21.42 |
| 3DRN [32]| 7.15 | 8.57 | 8.59 | 40.26 | 21.81 | 20.14 |
| GSTEG [16]| 7.52 | 9.08 | 9.08 | 40.86 | 22.42 | 20.31 |
| ST3DCNN| 7.84 | 9.46 | 9.87 | 41.76 | 22.30 | 21.68 |

A. DATASETS

In this section, we introduce two widely-applied public datasets, (ImageNet-VidVRD [14], VidOR [17]) and our new dataset (VidPDR) in the video visual relation detection.

ImageNet-VidVRD is the first evaluation dataset for VidVRD which consists of carefully selected videos from ILSVRC2016-VID. The dataset contains 135 categories of objects and 35 categories of predicates. Each video has many visual relations with clear annotations. In our experiments, we randomly selected 800 videos and 200 videos from ImageNet-VidVRD as the training set and the testing set. The testing set was annotated by 4,835 visual relation instances (video-level) in 1,011 triplet visual categories.

VidOR is a large-scale dataset composed of 10,000 videos with plentiful annotations on 80 categories of objects and 50 categories of predicates. The ACM MM’ 19 Video Relation Understanding (VRU) Challenge treats the VidOR as the public dataset. The data resources of VidOR come from YFCC-100M, which contains 0.8 million free and public multimedia videos. All the contents of videos are taken from the various scenarios in real life (e.g., “indoor”, “outdoor”, “entertainment”, “working scene”, etc.). The actual length of these videos varies from a few seconds to a few minutes and the average length is around 30 seconds. Compared with ImageNet-VidVRD, the average length (35.73s) of video in VidOR is longer for the reason that VidOR involves more complex real-life scenarios which makes it difficult to apply this dataset in video visual relation detection. For VidOR, we randomly selected 7,000 videos, 2,165 videos, and 835 videos as the training set, validation set, and testing set. Furthermore, the testing set is private for the grand challenge.

We contribute a novel video dataset, called Video Predicate Detection and Reasoning (VidPDR), consisting of 1000 videos with plentiful manually labeled annotations on 21 object classes and 37 predicate classes. In order to be closer to the actual scenario, we added some comparison predicate classes such as taller-than and shorter-than, which are more practical in some fields.

We collected the videos for VidPDR from the TrackingNet [36], a large-scale object tracking dataset containing 30,643 videos from Youtibe-BoundingBoxes. We removed all the videos that describe some limiting cases and are hard to be annotated. The filtering rules are as follows:

- resolution of the videos is too low to view clearly
- camera shakes during video leading to a blurred screen.
- only one object appears in the video
- too many objects in the video

Finally, we chose 1,000 high qualified videos from the TrackingNet to construct our dataset. We split the dataset into 800 videos for training and 200 videos for testing. We carefully annotated objects and predicates for each video in the whole dataset.

The object categories we use to annotate videos are composed of data in popular object detection datasets in the industry. To capture more fine-grained character relationships, we divided the categories related to characters in these datasets [37], [38] into person and children. The results of object categories are shown in Figure 4.

According to the statistics, the total number of annotated objects in the entire training set is 2,057. On average, there will be 2.1 objects appear in each video. As shown in Figure 4, we can observe that the number of objects in the entire video dataset roughly obeys the heavy-tailed distribution. More than half of the objects in videos are human.

As shown in Figure 5, 37 categories are defined as basic categories. These categories consist of 22 spatial predicates, 9 atomic action predicates, and 6 comparison predicates. It can be seen that our dataset has richer predicate categories than previous datasets.

Due to the ambiguity of the viewpoint in the spatial relationship, we normalize the viewpoint which comes from the objects instead of the camera. In general, the motion of the object in the video contains more information than the motion of the camera.

For example, the predicate “person-stand_behind-bicycle” indicates that a person is standing behind the bicycle no matter where the camera is. Another example is not available. The orientation of many regular-shaped objects is very difficult to determine when there is no reference (e.g. ball). Therefore, this kind of predicate cannot be used in the experiment. We calculated the number of relations based on the predicates in the training set. The statistics are shown in Figure 5. We can observe that among all the relations that appear in the video, the number of spatial predicates is the largest, followed by the action predicates, and finally the number of comparative predicates is the smallest. The number of such predicates is roughly the same as the number of spatial relations of the majority. This distribution is also in line with our experimental expectations and leads the model to learn more spatial relations.
In addition to studying the visual relationship at the predicate level, the model we proposed can also learn the visual relationship of triples level. Except for some situations that would not appear in real life, 14,000 triples are generated by combining predicates and relational pairs. The training set contains 1,258 categories of relation triplets (dark and grey area) and the testing set contains 510 categories of relation triplets. Moreover, the number of triples that exist in the whole dataset is 1,015, and the number that only appears in the validation set is 2. Some previous studies have shown that the distribution of this data makes the model unable to see the labels of the training stage in the prediction, thereby enhancing the generalization ability of the model.

As mentioned above, we chose two public datasets, ImageNet-VidVRD and VidOR, in the research field of video visual relation detection. And we contribute a new dataset named VidPDR which consists of 1,000 videos with dense manually labeled annotations. In our experiments, we verify our method performance comparison on the three datasets.

### B. IMPLEMENT DETAILS

Our proposed framework includes a three-stage training process. The object detection was first trained on VidPDR and achieved suitable performance. We use Intersection over Union (IoU) as loss function to optimize training parameters. Batchsize is set to 10 and Learning rate is set to 1e-3. Then, we used 3DCNN as the feature extraction model for relation prediction. MSELoss was used for this model and batchsize is set to 128 and the initial learning rate is set to 1e-3. We randomly divided the training dataset into 10 subsets and perform 10-fold cross-validation. Each validation is repeated...
10 times to achieve better performance. Our experiments are conducted with two Tesla V100-PCIe based on Pytorch.

C. EVALUATION METRICS
In our experiment, Recall@50, Recall@100, and mAP (mean Average Precision) are applied to verify the effectiveness. Recall@K denotes the fraction of correct video visual relation instances detected in the top K detection results and mAP is defined as:

$$mAP = \frac{\sum_{triplet \in C} AP}{|C|}$$

C is the number of relation triplet in testing set. AP means the average precision that indicates the number of correct videos.

D. COMPARED METHODS
To evaluate the effectiveness of the features extracted by the 3D Convolutional Neural Network, we compared our method with several video visual relation detection models.

1) VidVRD [14]
Shang et al proposed the first method called VidVRD, which uses a bottom-up strategy and divides the video into segments and predicts the video visual relationship between the common short-term object trajectories. After that, a complete relationship instance is generated through the greedy association algorithm. This paragraph simplifies the introduction of the method.

2) MFF [22]
Sun et al adopt the same strategy with VidVRD, and it explicitly combines time and space features and textual information extracted from all segments for prediction. In our experiments, the method is MFF. We need to extract the relative location features and moving features. The relative location feature, $f_{Loc} = [s_x, s_y, s_w, s_h, s_d]$, is a widely applied to encode spatial feature [39]. The proposed moving feature describes the location information between subject and object:

$$f_{Mot} = f^s_{Loc} - f^e_{Loc}$$

where $f^s_{Loc}$ and $f^e_{Loc}$ are the location feature from start frame and the location feature from the end frame, respectively. The spatial-temporal feature $f_{ST}$ is generated by aggregate the features mentioned above as $[f^s_{Loc}, f^e_{Loc}, f_{Mot}]$. The context feature $f_{Chan}$ is a representation with 1,200 dimensions, which is generated by concatenating the embedding of subject and object.

3) 3DRN [32]
Existing visual relation methods were divided into sub-task. Each sub-task generated different features which lack the information flows between them. Cao et al proposed 3DRN that can connect the separated representations to share a set of task-specific features.

4) GSTEG [16]
Tsai et al proposed a Conditional Random Field on the fully-connected graph to model the relationship between different objects spatially and temporally. A parametrization based on a gated energy function was introduced to learn adaptive features of frames.

5) VidVRD+MHA [33]
Compared to relation detection on images, visual relation association plays an important role in video visual relation detection. To release the effect of inaccurate tracklet detection and relation prediction, Su et al proposed a relation association method, called Multiple Hypothesis Association (MHA), to improve the performance of VidVRD.

6) VRD-GCN [15]
Qian et al proposed a graph structure based method named VRD-GCN to achieve better results of prediction on objects as well as their dynamic relationships by taking advantage of spatial-temporal contextual information.

7) RELAbuilder [34]
Traditional video visual detection pipeline consists of many complex stages. Therefore, Zheng et al followed the basic framework which included object detection, object tracker, and relation predictor. Researches also proposed RELAbuilder to solve the data unbalance and label missing problems.

8) MAGUS.Gamma [22]
Sun et al proposed a video visual relation detection method with multi-model feature fusion, which earned the first place in visual relation detection task of Video Relation Understanding Challenge (VRU), ACM MMM 2019.

9) VRD-STGC [35]
Dividing one video into several video segments and merging predict relationships is a common method in video visual relation detection tasks. However, such method cannot capture the relationship involving long time. Liu et al introduce a sliding-window scheme to predict the long-term relationship.

V. RESULT
As shown in Table 1,2,3, we report the result of compared models on two public datasets (VidOR and ImageNet-VidVRD) and self-constructed datset (VidPDR). In the hyperparameters setting, most of the parameters were set by following the original papers. From these tables, we can have some observations. Table 1,2, shows the performances of the baseline model and our method on classical public datasets (ImageNet-VidVRD and VidOR).

1) the metrics of the ST3DCNN, VRD-STGC, VRD-GCN, MFF, and 3DRN has a better performance than other baselines.
2) ST3DCNN acquires 2.07%, 0.46%, and 14.29% relatively higher R@50, mAP, and P@1 than VidVRD on ImageNet-VidVRD.

3) ST3DCNN acquires 2.72%, 5.88%, and 6.79% relatively higher R@100, P@5, and P@10 than VidVRD on VidOR.

4) Compared to VRD-GCN which apply the normal graph neural network, ST3DCNN achieves that the relative improvements of R@100, P@5, and P@10 are 0.21%, 0.13%, and 0.9%, respectively.

Table 3, shows the performances of methods on VidPDR. As we can see, ST3DCNN also gets good results. ST3DCNN which proves that the features extracted by 3D Convolutional Neural Network perform better than some manual encoded features in the video visual relation detection achieved best performance on most of evaluation metrics. In addition, it can be seen from the above analysis that ST3DCNN is more precise than other models, which can be used in scenes needed more precision. In the future research, we can study more in the direction of better extraction of visual features.

VI. CONCLUSION

In our work, we proposed a new dataset for the task of video visual relation detection, which is named VidPDR consisting of 1,000 videos with dense manually dynamic labeled annotations on 21 categories of objects and 37 categories of predicates. Furthermore, we introduce a new method that could extract some more abundant and comprehensive Spatio-temporal features by the 3D Convolutional Neural Network. Experiments on three datasets show that the Spatio-temporal features extracted by 3D Convolutional Neural Network could increase the performance of video visual relation detection.

REFERENCES

[1] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, “Heterogeneous memory enhanced multimodal attention model for video question answering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). Jun. 2019, pp. 1999–2007.

[2] J. Gao, R. Ge, K. Chen, and R. Nevatia, “Motion-appearance co-memory networks for video question answering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). Jun. 2018, pp. 6576–6585.

[3] S. Liu, Z. Ren, and J. Yuan, “Si2Net: Sibling convolutional encoder for video captioning,” in Proc. 26th ACM Int. Conf. Multimedia, Oct. 2018, pp. 1425–1434.

[4] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, “Streamlined dense video captioning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). Jun. 2019, pp. 6588–6597.

[5] J. Xu, Y. Yao, Y. Zhang, and T. Mei, “Learning multimodal attention LSTM networks for video captioning,” in Proc. 25th ACM Int. Conf. Multimedia, Oct. 2017, pp. 537–545.

[6] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, “Real-time video super-resolution with spatio-temporal networks and motion compensation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). Jul. 2017, pp. 2848–2857.

[7] M. Rochan and Y. Wang, “Video summarization by learning from unpaired data,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). Jun. 2019, pp. 7902–7911.

[8] B. Zhao, X. Li, and X. Lu, “HSA-RNN: Hierarchical structure-adaptive RNN for video summarization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jan. 2018, pp. 7405–7414.

[9] M. Douze, J. Revaud, J. Verbeek, H. Jégou, and C. Schmid, “Circulant temporal encoding for video retrieval and temporal alignment,” Int. J. Comput. Vis., vol. 119, no. 3, pp. 291–306, Sep. 2016.

[10] K. Schoeffmann, M. J. Primus, B. Münzer, S. Petschnigg, C. Kirisch, Q. Xu, and W. Hurst, “Collaborative feature maps for interactive video search,” in Proc. Int. Conf. Multimedia Modeling, 2017, pp. 457–462.

[11] J. Song, L. Gao, L. Liu, X. Zha, and N. See, “Quantization-based hashing: A general framework for scalable image and video retrieval,” Pattern Recognit., vol. 75, pp. 175–187, Mar. 2018.

[12] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” Int. J. Comput. Vis., vol. 123, no. 1, pp. 32–73, May 2017.

[13] C. Lu, R. Krishna, M. S. Bernstein, and F. Li, “Visual relationship detection with language priors,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 852–869.

[14] X. Shang, T. Ren, J. Guo, H. Zhang, and T.-S. Chua, “Video visual relation detection,” in Proc. 25th ACM Int. Conf. Multimedia, Oct. 2017, pp. 1300–1308.

[15] 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, Oct. 2019, pp. 84–93.

[16] Y.-H. Tsai, S. Divvala, L.-P. Morency, R. Salakhutdinov, and A. Farhadi, “Video relationship reasoning using gated spatio-temporal energy graph,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). Jun. 2019, pp. 10424–10433.

[17] X. Shang, D. Di, J. Xiao, Y. Cao, X. Yang, and T.-S. Chua, “Annotating objects and relations in user-generated videos,” in Proc. Int. Conf. Multimedia Retr., Jun. 2019, pp. 279–287.

[18] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inform. Process. Syst., 2015, pp. 91–99.

[19] Y. Li, W. Ouyang, X. Wang, and X. Tang, “VIP-CNN: Visual phrase guided convolutional neural network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 7244–7253.

[20] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. Elgammal, “Relationship proposal networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5226–5234.

[21] S. D. Khan, L. Alarabi, and S. Basalamah, “A unified deep learning framework of multi-scale detectors for geo-spatial object detection in high-resolution satellite images,” Arabian J. Sci. Eng., pp. 1–16, Oct. 2021.

[22] X. Sun, T. Ren, Y. Zi, and G. Wu, “Video visual relation detection via multi-modal feature fusion,” in Proc. 27th ACM Int. Conf. Multimedia, Oct. 2019, pp. 2657–2661.

[23] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, “Flow-guided feature aggregation for video object detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 408–417.

[24] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang, “Seq-NMS for video object detection.” 2016, arXiv:1602.08465.

[25] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-speed tracking with kernelized correlation filters,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 583–596, Mar. 2015.

[26] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, Jan. 2013.

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 5534–5542.

[28] I. Carreira and A. Zisserman, “Quo vadis, action recognition? A new view on temporal modeling,” in Proc. 27th ACM Int. Conf. Multimedia, Oct. 2017, pp. 5534–5542.

[29] J. M. Douze, J. Revaud, J. Verbeek, H. Jégou, and C. Schmid, “Circulant temporal encoding for video retrieval and temporal alignment,” Int. J. Comput. Vis., vol. 119, no. 3, pp. 291–306, Sep. 2016.

[30] K. Schoeffmann, M. J. Primus, B. Münzer, S. Petschnigg, C. Kirisch, Q. Xu, and W. Hurst, “Collaborative feature maps for interactive video search,” in Proc. Int. Conf. Multimedia Modeling, 2017, pp. 457–462.
[32] Q. Cao, H. Huang, X. Shang, B. Wang, and T.-S. Chua, “3-D relation network for visual relation recognition in videos,” Neurocomputing, vol. 432, pp. 91–100, Apr. 2021.
[33] Z. Su, X. Shang, J. Chen, Y.-G. Jiang, Z. Qiu, and T.-S. Chua, “Video relation detection via multiple hypothesis association,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 3127–3135.
[34] S. Zheng, X. Chen, S. Chen, and Q. Jin, “Relation understanding in videos,” in Proc. 27th ACM Int. Conf. Multimedia, Oct. 2019, pp. 2662–2666.
[35] C. Liu, Y. Jin, K. Xu, G. Gong, and Y. Mu, “Beyond short-term snippet: Video relation detection with spatio-temporal global context,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10840–10849.
[36] M. Müller, A. Bibi, S. Giancola, S. Al-Subaihi, and B. Ghanem, “TrackingNet: A large-scale dataset and benchmark for object tracking in the wild,” in Proc. Eur. Conf. Comput. Vis., Sep. 2018, pp. 310–327.
[37] 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 Proc. Eur. Conf. Comput. Vis., Sep. 2014, pp. 740–755.
[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. 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, Dec. 2015.
[39] 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), Jul. 2017, pp. 3107–3115.

MINGCHENG QU received the Ph.D. degree in computer science and technology from the Harbin Institute of Technology, in 2011. He is currently a Lecturer and a Master Supervisor with the Faculty of Computing, Harbin Institute of Technology. His current research interests include artificial intelligence, computer vision, distributed and big data computing, embedded/thef Internet of Things systems, and robot application technology. More than 50 papers have been published or accepted by domestic and foreign important academic journals and conferences and more than 30 papers have been indexed by SCI/EI/CPCI. Single paper has been downloaded more than 3000 times. He has undertaken multiple projects, including the National Natural Science Foundation of China (61402131), China Postdoctoral Science Foundation Special Fund (2016T90293), and China Postdoctoral Science Foundation (2014M551245).

JIANXUN CUI is currently an Associate Professor with the School of Transportation Science and Engineering, Harbin Institute of Technology. He is mainly engaged in the interdisciplinary research of deep learning and intelligent transportation. He has presided over the National Natural Science Foundation of China (NSFC) Youth Program, the National Natural Science Foundation of China (NSFC) Emergency Management Program, the 863 Project Sub-Project, and the China Postdoctoral Special Funded Program. He has published more than ten SCI/SSCI indexed articles, eight EI indexed articles, and three academic monographs. Apply for ten authorized invention patents.

TONGHUA SU received the Ph.D. degree in computer science and technology from the Harbin Institute of Technology, in 2008. He is currently an Associate Professor and a Master Supervisor with the Faculty of Computing, Harbin Institute of Technology. He has published more than 20 research papers in reputed journals and conferences, such as ICFR, ICFHR, ICDAR, and CCPR. His research interests include pattern recognition and deep learning, the Internet of Things big data intelligent information processing, and intelligent media computing and interaction technology. He is a Reviewer of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), Pattern Recognition, Information Sciences, and International Journal on Document Analysis and Recognition.

GANLIN DENG is currently pursuing the degree in computer science and technology with the Harbin Institute of Technology. Her research interests include deep learning, computer vision, and video caption.

WENKAI SHAO received the B.E. degree from Xidian University, Xi’an, China, in 2018. He is currently pursuing the M.A. (Eng.) degree with the Faculty of Computing, Harbin Institute of Technology. His research interests include deep learning, computer vision, and visual content understanding. * * *