Semantic-Guided Multi-feature Fusion for Accurate Video Captioning

Yunjie Zhang[1], Tianyang Xu[1(✉)], Xiaoning Song[1], Zhenghua Feng[2], and Xiao-Jun Wu[1]

[1] Jiangnan University, Wuxi 214122, China
6213113148@stu.jiangnan.edu.cn,{tianyang.xu,x.song,wu_xiaojun}@jiangnan.edu.cn
[2] University of Surrey, Guildford GU2 7XH, UK
z.feng@surrey.ac.uk

Abstract. In recent years, pre-trained visual language models (PVLMs) have achieved superior performance in many downstream tasks by extracting comprehensive cross-modal relevance from billions of pieces of data. Video captioning is a typical topic that aims to generate semantic texts from video clips, which also benefits from the advances in PVLMs. However, existing PVLMs only extract holistic features from still images, neglecting the local and temporal changes in the video appearance, which impedes fine-grained video understanding. Drawing on this, we propose to add explicit spatio-temporal semantics to the existing video captioning system by wrapping the detected salient objects over sampled frames, reflecting thematic events within a video. In particular, an auxiliary detection branch is designed to collaborate with PVLMs, achieving fine-grained object awareness. To achieve efficient temporal aggregation, we further employ the Gated Recurrent Unit (GRU) to extract temporally ordered cues, compensating for the limited temporal appearance capacity of PVLMs. The experimental results obtained on several benchmark datasets demonstrate the effectiveness of the proposed solution, with superior performance compared to the state-of-the-art approaches.

Keywords: Video Captioning · Object Detection · Gated Recurrent Unit

1 Introduction

Video captioning aims to understand the events in a video, with the ability to automatically predict captions, which has many practical applications in pattern recognition and computer vision, *e.g.*, video summary, video key detection,

This work is supported in part by the National Natural Science Foundation of China (Grant No.62106089, 62020106012); National Social Science Fundation of China(21&ZD166); Natural Science Foundation of Jiangsu Province, China (BK20221535).

H. Lu et al. (Eds.): ICIG 2023, LNCS 14358, pp. 47–58, 2023.
https://doi.org/10.1007/978-3-031-46314-3_4

and blind navigation. A traditional video captioning framework typically first extracts hand-crafted visual features from a given video clip. After obtaining feature representations, the video subtitle generation system generates sentences using predefined templates. The effectiveness of this framework is highly dependent on the predefined templates, while fixed templates always result in fixed syntactic structures in the generated caption sentences [16]. In recent years, deep learning has become the mainstream technique for video captioning, similar to the development of other visual or language tasks. In general, many sequential learning networks adopt the encoder-decoder architecture to flexibly generate caption output. The seminal encoder-decoder model is Sequence to Sequence Video to Text (S2VT) [9] which has two stacked Long Short Term Memories (LSTMs).

S2VT is the initial deep learning-based framework for the video description task and it is the first to introduce the encoder-decoder structure to the video description task, which has inspired many subsequent models. For example, Spatio-Temporal Attention Long Short Term Memory (SA-LSTM) [10] presents an attention method that combines local and global temporal structures of video features and considers different motion patterns, which can effectively generate accurate video captions. Reconstruction Network (RecNet) [24], on the other hand, is a novel video caption reconstruction network, which not only adopts video-to-text generation but also explores text-to-video mapping, unifying the semantic space between the two modalities [5].

Almost all of the above approaches use Convolutional Neural Networks as video encoders. However, since CNNs are experts in processing visual features, but lack the power of textual semantic extraction, it is not optimal to implement video captioning using pure CNNs. To solve this problem, Transformer methods [1], which construct inter-dependencies between vision and language, are proposed to provide improved generation capabilities over CNN approaches. Nevertheless, the task is still far from being solved due to the inconsistency between video appearance and language cues.

The PVLM has received considerable attention in recent years for bridging the complex semantic gap between images and texts. It learns large and sophisticated patterns in a Transformer-like network and has achieved superior performance in a number of well-known benchmarks [2,3] and competitions. In principle, the success of PVLM lies in its transferability to the downstream tasks in terms of preserving the intrinsic discrimination and perception between the two modalities involved. Despite its power in bridging images and texts, PVLM is not tailored for video captioning, as it focuses on extracting holistic features from still images and text [19]. Therefore, the essential issues that need to be addressed for PVLM-based video captioning are two-fold: firstly, how to explore the local spatial semantics rather than holistic representations; secondly, how to perceive temporally ordered appearance variations in videos.

To mitigate the first issue, additional spatial modelling techniques have been studied accordingly. Yu *et al.* [20] propose to use a spatial attention mechanism to focus on local spatial semantics. However, this approach performs poorly

when detecting overlapping objects. With the development of target detectors, some methods attempt to extract local spatial semantics using target detectors. For instance, Zheng et al. [14] use a target detector to detect multiple object targets and further focus on the interactions between targets to generate high-quality predicates and verb subtitles. Therefore, we inherit the previous research methodology, using powerful object detection modules to explore the comprehensive capacity between holistic and local spatial semantics. To address the second issue, various designs have been proposed to reflect temporal cues from input videos. Cho *et al.* [4] use the GRU model to obtain a temporal representation of the cross-frame motion patterns. Zhang *et al.* [21] further use bi-directional temporal maps to capture the temporal trajectory of each salient object as a way to obtain motion relevance cues between video frames. Considering the absence of temporal representation within video features, in our work, we utilise a lightweight Gated Recurrent Unit (GRU) to perceive temporally ordered appearance variations in videos.

To summarise, the main contributions of the proposed method include:

- Local semantics are highlighted by an object detection module. This provides complementary visual cues for accurate video captioning. A dedicated multi-feature fusion module is employed to balance the saliency between object semantics and scenario overview.
- Temporally ordered cues are moderately aggregated via GRU, eliminating information redundancy among video frames. The temporal order can also be reflected by GRU, highlighting the potential causality of video data.
- State-of-the-art results are obtained on the MSVD [2] and MSR-VTT [3] datasets, demonstrating the effectiveness and robustness of the proposed approach.

2 Related Work

Video Captioning Based on CNN and RNN. In the early days, there are many traditional approaches used to formulate the video captioning task. In the beginning, Kojima *et al.* [8] propose a template-based method that predicts the words represented by specific objects and actions in video frames. Although straightforward, this approach suffers from the obvious disadvantage of not being able to generate diverse and flexible video descriptions. To alleviate this limitation, encoder-decoder architecture is adopted for video captioning to simultaneously predict the sequential output. Venugopalan *et al.* [9] are the first to explore the encoder-decoder structure for video captioning. They use CNN to extract video features from each frame and perform pooling operations to obtain a global video representation, and then generate the output captions with an LSTM module. Although this structure can extract descriptive visual features, it cannot interact visual features with textual features, so it lacks support from textual semantics. To remedy this shortcoming, Transformer-based methods are now widely used.

Transformer Techniques in Video Captioning. The success of Transformer models in natural language processing tasks has been transferred to the computer vision field in recent years. Due to the powerful attention mechanism, Transformer architectures are also widely used in the video captioning field. As LSTM cannot address the long-term dependency in the process of video encoding, Zhou *et al.* [1] proposed to use a Transformer instead of LSTM to extract relevant video features. Furthermore, to highlight the multi-modal property, Ging *et al.* [17] proposed a multi-layer Transformer structure that facilitates the semantic alignment of visual and textual features in a common embedding space. The Transformer-based encoder-decoder structure dominates the current designs, which is also our baseline structure. However, fine-tuning a Transformer model in the training stage often requires huge computational expenses, which impedes its practical applications. In order to reduce the computational burden, the use of PVLMs has become the most popular method, with promising transferability to downstream multi-modal tasks.

Pre-trained Visual-Language Models. PVLMs establish powerful multi-modal interactions by training on large-volume image-text pairs, bridging the semantic gap between the vision and language data. In particular, the CLIP model proposed by Radford *et al.* uses contrast learning to perform unsupervised training of images with massive texts. For downstream extensions, Li *et al.* propose ALBEF [18], which uses a detector-free image encoder and a text encoder to encode images and text independently. Specifically, we use the CLIP4Clip model to extract high-performance visual representations. Although PVLMs provide relevant connections between visual and language pairs, they tend to focus on holistic spatial semantics at the expense of neglecting local spatial semantics [23]. Therefore, it is necessary to explore the target local semantics in order to obtain enhanced perceptions that can promote accuracy and concentration during video-text alignment.

Utilising Local Semantics. Local target details play an important role in generating high-quality headlines. In order to perceive local semantics, pre-trained object detection models, *e.g.*, YOLO [12] and Faster-RCNN [6], have been widely studied in general computer vision field. In terms of video captioning, there have also been existing attempts to exploit the detected semantic information. In the work proposed by Aafaq *et al.* [13], the pre-trained YOLO object detector is used to extract the locations and scales of objects. Similarly, Ye *et al.* [22] used a pre-trained Faster-RCNN object detector to extract salient objects, with a multi-level modular network being constructed to effectively analyse the relationship among these objects, delivering accurate video captioning [26]. Consistent with the above development, we also aim to balance the local and holistic semantics of individual detection and PVLMs, respectively.

3 The Proposed Approach

In order to introduce local semantics and temporal information into the current captioning model, our model uses a target detector and a GRU model to extract local spatial semantics and temporally ordered cues, respectively. Then, the fusion module is used to fuse the holistic spatial pattern, local semantic information, and temporal clue. Finally, the Symmetric Cross Entropy (SCE) Loss [7] is used to guide the training of the model. The details of the overall structure of the method are shown in Fig. 1.

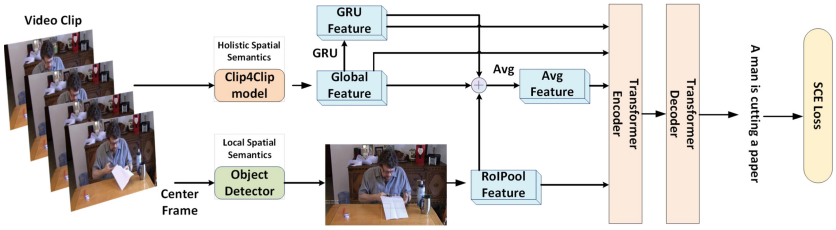

Fig. 1. Illustration of the proposed network structure that integrates multi-frame global features and local detection cues, compensating for the incomplete perception of local and holistic semantics. The fused features are input to the encoder-decoder Transformer blocks with Symmetric Cross Entropy (SCE) Loss for video captioning.

3.1 Spatial Semantics

To reflect comprehensive spatial appearance, both the holistic and local semantics of a video should be explicitly represented. In our method, twelve frames are extracted from each video sequence. The holistic spatial semantics of all twelve frames is obtained by the powerful CLIP4Clip model [25] with freezing parameters, while the object detection module is applied for extracting the local spatial semantics from the centre frame. In particular, the object detection module is used to store the locations of local entities in the video frame. Then the features are extracted from these local spatial regions using the convolutional layers of ResNet50. Finally, a Multi-layer Perceptron (MLP) module is used to further strengthen the features and unify the dimensions, so that they can be projected into the same feature space with other features.

In view of the superior universality, robustness and better performance of the Faster-RCNN object detector, we use it as our target detector. In this paper, the Faster-RCNN object detector with freezing parameters is used to detect the objects in each centre frame. By using the Faster-RCNN, the local spatial semantics obtained is $N \times 1024$, where N is the number of detected targets, 1024 is the candidate feature dimension on the convolutional layer output. Since we prefer to obtain more local spatial semantics, the classification threshold of Faster-RCNN is set to 0.3. Using a lower classification threshold allows the detector to obtain more objects in a video frame. Even if the lower classification

threshold results in inaccurate classification results, we can still obtain effective spatial local semantics. Because we value the spatial local semantics within the prediction box, we focus on the accuracy of the prediction box rather than the accuracy of the classification results. By paying attention to both local and holistic spatial semantics, a more comprehensive representation of video features can be obtained.

3.2 Temporally Ordered Representations

Despite the holistic and local spatial semantics provided by the above appearance model, temporally ordered cues are currently neglected. The absence of time-ordered cues may lead to unclear changes of action between objects over time. To achieve efficient temporal aggregation, we further employ a GRU module to emphasise the temporal relevance of text-related holistic spatial semantics extracted from the Clip4Clip model. Since we train with features extracted by PVLM, we believe that our features have sufficiently learned the video representation, so that additional complex models are not necessary to obtain the temporal representation. Therefore, we use GRU to obtain temporal cues in features, taking into account the parsimony of the GRU structure.

In this module, the size of the holistic spatial semantics is set to 12×512, where 12 is the number of frames and 512 is the feature dimension. To unify the feature dimensions and extract the inter-frame relationship from the 12 video frames, the size of the temporally ordered representations is 1×512 after the GRU module. Effective temporal series representation can compensate for the lack of temporal cues in the features and enrich the video feature representation.

3.3 Feature Fusion Method

After obtaining the holistic spatial semantics, the local spatial semantics, and the temporally ordered representations, it is essential to effectively integrate these features.

In our design, since the size of the local spatial semantics is $N \times 1024$, which does not match the dimension of the holistic spatial semantics, a linear projection layer is used to reduce its dimensionality to $N \times 512$. Then, we cascade the holistic spatial semantics, the local spatial semantics, and the temporally ordered representations in the feature dimension to obtain the fused features. The plus sign indicates that the individual features are combined by concatenation. The characteristic dimension of the fusion is $(13+N) \times 512$. Where N is the number of objects detected by the target detector. The 13 dimensions contain 12 dimensional frame features and 1 dimensional temporally ordered features.

Next, the global average pooling is applied to obtain the global average features with the size of 1×512. Last, the global average features are merged with the fused features, and the final size of the Transformer input is $(14 + N) \times 512$. The 14 dimensions contain 12 dimensional frame features, 1 dimensional global average features, and 1 dimensional temporally ordered features.

The fusion of multiple feature representations as described above allows for a comprehensive consideration of global and detailed features, holistic and partial features, and the incorporation of temporal cues. The complementarity between multiple features is exploited to obtain a video representation that is more semantic, less noisy and contains more critical information. To make a long story short, good features are the key to improving the effectiveness of the model.

3.4 The SCE Loss

Since the video labels are generally noisy and blurred, we use the SCE Loss instead of the original Cross Entropy (CE) Loss to relieve over-fitting and against noise with a regular term. The specific approach is to use the SCE Loss to relax the original strict binary label. We slightly decrease the value of the correct label from 1 and increase the values of the other categories from 0 to relax the strict constraint of cross entropy. SCE Loss is a combination of Cross Entropy (CE) Loss and Reverse Cross Entropy (RCE) Loss. CE Loss and RCE Loss are defined as follows:

$$\mathbf{L}_{ce} = -\sum_{t=1}^{L} P(t) \log Q(t) \tag{1}$$

$$\mathbf{L}_{rce} = -\sum_{t=1}^{L} Q(t) \log P(t) \tag{2}$$

where P and Q are the predictions and real outputs respectively. $\mathbf{L}_{ce}$ is the normal cross entropy loss, $\mathbf{L}_{rce}$ is cross entropy loss of switched labels. The SCE Loss is defined as:

$$\mathbf{L}_{sl} = \lambda_1 \mathbf{L}_{ce} + \lambda_2 \mathbf{L}_{rce}, \tag{3}$$

where λ_1 and λ_2 are two hyper-parameters.

By smoothing the labels in this way, we relax the original strict classification prediction results, so that the predicted captions can be some synonyms of the ground truth captions, improving the universality and rationality of the predicted captions.

4 Experimental Results

We evaluate the proposed method on two publicly available data sets, *i.e.*, MSVD and MSR-VTT. The used evaluation metrics of MSVD and MSR-VTT are BLUE@4(B@4), METEOR(M), ROUGE-L(R) and CIDEr(C).

Table 1. Ablation studies on MSVD and MSR-VTT

	MSVD					MSR-VTT				
Methods	B@4↑	M↑	R↑	C↑	Params↓	B@4	M	R	C	Params
Baseline	57.1	40.0	76.8	114.0	81MB	46.8	31.3	64.8	60.1	81MB
Baseline+OD	58.7	40.8	77.7	117.9	81MB	48.0	31.7	65.2	60.7	81MB
Baseline+OD+GRU	59.1	41.0	77.6	119.4	85MB	48.4	31.7	65.5	61.1	85MB

4.1 Ablation Study

In order to verify the effectiveness of the proposed method, we first report the corresponding ablation analysis. Table 1 reports the performance on MSVD and MSV-VTT datasets. In general, OD represents the object detection module, GRU represents the GRU module. Our baseline is the CLIP4Clip model equipped with holistic spatial semantics. We then test the baseline model with local spatial semantic features. Finally, we experiment with models with additional temporal sequence cues.

The impact of the OD module. Compared to the baseline, the use of local spatial semantics increases the performance in terms of CIDEr by 3.9 and 0.6 on the two data sets. The exploration of local spatial semantics can compensate for the shortcomings of PVLM in extracting only holistic features from images, which is the main reason for the improvement of our evaluation metrics.

The impact of the GRU module. By integrating the temporally ordered cues through GRU, we can further improve the performance by 5.4 and 1.0 in terms of CIDEr on the two datasets. Due to the particularity of the video task, the temporal feature transformation in the video is extremely important compared to the static image feature. Therefore, the GRU module can be used to sense the appearance change of the temporal sequence in the video, which can further improve the evaluation index.

In addition, although there are more variables involved in our design, the increase in parameters is less than 10%. This is mainly due to the fact that the lightweight GRU module we use does not increase the number of parameters by a large amount. Given that our baseline is already able to extract valid global spatial transformations, it is perfectly adequate to use the lightweight GRU module to compensate for the lack of temporal cues.

We also test two methods to reduce the dimension of local spatial semantics, Linear and Transformer, respectively. For transformer, we use a layer of transformer encoder structure and change the size of its output dimensions. According to the experimental results in Table 2, leading results can be obtained by directly using linear projection. Based on this, we believe that the Transformer, which focuses on strong relationships between features, can support improved visual semantics. However, the loss of superficial feature details undermines the valid feature relationships extracted in PVLM. Based on this, the use of the Transformer to convert dimensions yields poor results in this paper.

Table 2. Different methods of dimensionality reduction on MSVD

Methods	B@4	M	R	C
Linear	58.7	40.8	77.7	117.9
Transformer	53.8	38.8	75.5	104.1

4.2 Comparison to State-of-The-Art

To demonstrate our modelling merits, we compare the proposed method with 15 state-of-the-art approaches on MSVD and MSR-VTT benchmarks. The results are reported in Table 3. As can be seen from the table, on the MSVD dataset, the performance of the proposed method is only 0.1 lower than that of HMN in the BLEU@4 evaluation index, while the performance of METEOR, ROUGE-L and CIDEr is higher than that of the previous optimal methods respectively. On the MSR-VTT dataset, our method outperforms all other methods on all evaluation indices.

This is mainly due to the fact that we exploit the sufficient prior knowledge in PVLM and the use of the target detector and the temporal model to compensate for the lack of local semantic features and temporal cues in PVLM. By fully integrating multi-scale and multi-angle features, our method is more comprehensive and versatile. Its excellent performance in each evaluation index also confirms the advantages and superiority of the method over other approaches.

Table 3. Comparison with 15 state-of-the-art MSVD and MSR-VTT benchmark methods. The best results are shown in bold.

				MSVD				MSR-VTT			
Methods	Backbone	Features Motion	Object	B@4	M	R	C	B@4	M	R	C
M3 (CVPR-18)	VGG	C3D	-	51.8	32.5	-	-	38.1	26,6	-	-
RecNet (CVPR-18)	Inception-V4	-	-	52.3	34.1	69.8	80.3	39.1	26.6	59.3	42.7
PickNet (ECCV-18)	ResNet-152	-	-	52.3	33.3	69.6	76.5	41.3	27.7	59.8	44.1
MARN (CVPR-19)	ResNet-101	C3D	-	48.6	35.1	71.9	92.2	40.4	28.1	60.7	47.1
OA-BTG (CVPR-19)	ResNet-200	-	Mask-RCNN	56.9	36.2	-	90.6	41.4	28.2	-	46.9
POS-GG (ICCV-19)	InceptionResnetV2	OpticalFlow	-	52.5	34.1	71.3	88.7	42.0	28.2	61.6	48.7
MGSA (AAAI-19)	InceptionResnetV2	C3D	-	53.4	35.0	-	86.7	42.4	27.6	-	47.5
GRU-EVE (CVPR-19)	InceptionResnetV2	C3D	YOLO	47.9	35.0	71.5	78.1	38.3	28.4	60.7	48.1
STG-KD (CVPR-20)	ResNet-101	I3D	Faster-RCNN	52.2	36.9	73.9	93.0	40.5	28.3	60.9	47.1
SAAT (CVPR-20)	InceptionResnetV2	C3D	Faster-RCNN	46.5	33.5	69.4	81.0	40.5	28.2	60.9	49.1
ORG-TRL (CVPR-20)	InceptionResnetV2	C3D	Faster-RCNN	54.3	36.4	73.9	95.2	43.6	28.8	62.1	50.9
SGN (AAAI-21)	ResNet-101	C3D	-	52.8	35.5	72.9	94.3	40.8	28.3	60.8	49.5
MGRMP (ICCV-21)	InceptionResnetV2	C3D	-	55.8	36.9	74.5	98.5	41.7	28.9	62.1	51.4
HMN (CVPR-22)	InceptionResnetV2	C3D	Faster-RCNN	**59.2**	37.7	75.1	104.0	43.5	29.0	62.7	51.5
CLIP4Caption (CVPR-21)	CLIP4Clip		-	-	-	-	-	46.1	30.7	63.7	57.7
Ours	CLIP4Clip		Faster-RCNN	59.1	**41.0**	**77.6**	**119.4**	**48.4**	**31.7**	**65.5**	**61.1**

4.3 Qualitative Results

We present qualitative results in Fig. 2, from which we can see that the proposed method can generate high-quality captions. The old method often produced wrong subtitles, see Wrong in Fig. 2, and the wrong place is usually the

Fig. 2. Qualitative results on MSVD. The images are the sampled frames of two videos. The image on the right is the feature attention map of the features extracted from the CLIP4Clip model. The text "GT" represents the ground truth video captions, "Wrong" represents the wrong video captions generated by the baseline model, and "Ours" represents our generated captions respectively.

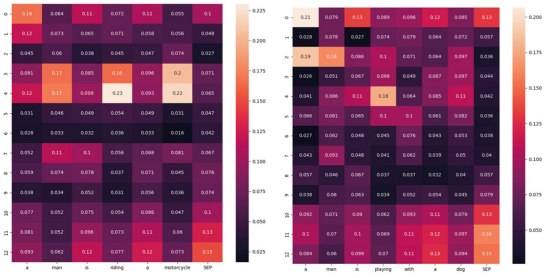

Fig. 3. The heat map relationship matrix between video captions and video features. The abscissa represents our predicted video captions, where SEP represents the end character of the predicted captions. The ordinate represents the video frame extracted from the video. If the text is closely related to the video features, the corresponding values are large and the colours are bright.

subject, object or verb. Therefore, in our approach, we use different models to further focus on area objects and the action relationships between objects. These practices often lead to more accurate subject, object or verb predictions. See Ours in Fig. 2. In addition, based on the attention heat maps obtained from our transformer decoder on the right side of Fig. 2, the region of salient objects can be correctly selected by the proposed method, indicating that the proposed method can distinguish the objects from their surroundings. More importantly, our model can also ignore some redundant frames. In the second example, the proposed method only focuses on the man and the motorbike, rather than the tire that appears in the first frame.

We also show the heat map relationship matrix between video captions and video features, as shown in Fig. 3. We choose two examples that match those shown in Fig. 2. As can be seen in Fig. 3, nouns and verbs are closely related to video features. This is mainly due to the addition of a target detection model and a GRU model of perceptual temporal cues to our approach, which focuses more on regional objects and actions. And, since not all video frames are relevant

to the caption, some are redundant and our method is able to focus on the key video frames and ignore the irrelevant ones. The closest relationship tends to focus on certain key frames.

5 Conclusion

This paper presents a semantic-guided multi-feature fusion approach for accurate and robust video captioning. The proposed method harmonises holistic spatial semantics, local spatial semantics, and temporally ordered representations for high-performance video captioning. By constructing an effective feature fusion method, the proposed approach fuses the above features via attention operations, to obtain comprehensive visual representations of captions. Meanwhile, the SCE Loss is advocated for training the Transformer model with relaxed supervision. The proposed method achieves state-of-the-art performance on both the MSVD and MSR-VTT benchmarks, validating the merits of the method.

References

1. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.,: End-to-end dense video captioning with masked transformer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8739–8748 (2018)
2. Guadarrama, S., Krishnamoorthy.: YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2712–2719 (2013)
3. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: a large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5288–5296 (2016)
4. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078(2014)
5. Xu, T., Wu, X.J., Kittler, J.: Non-negative subspace representation learning scheme for correlation filter based tracking. In: 2018 24th International Conference on Pattern Recognition, pp. 1888–1893 (2018)
6. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
7. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 322–330 (2019)
8. Kojima, A., Tamura, T., Fukunaga, K.: Natural language description of human activities from video images based on concept hierarchy of actions. Int. J. Comput. Vision **50**, 171–184 (2002). https://doi.org/10.1023/A:1020346032608
9. Venugopalan, S ., et al.: Sequence to sequence-video to text. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4534–4542 (2015)
10. Yao, L., et al.: Describing videos by exploiting temporal structure. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4507–4515 (2015)
11. Hori, C., et al.: Attention-based multimodal fusion for video description. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4193–4202 (2017)

12. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
13. Aafaq, N., Akhtar, N., Liu, W., Gilani, S.Z., Mian, A.: Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12487–12496 (2019)
14. Zheng, Q., Wang, C., Tao, D.: Syntax-aware action targeting for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13096–13105 (2020)
15. Pan, B., et al.: Spatio-temporal graph for video captioning with knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10870–10879 (2020)
16. Xu, T., Feng, Z.H., Wu, X.J., Kittler, J.: An accelerated correlation filter tracker. Pattern Recogn. **102**, 107172 (2020)
17. Ging, S., Zolfaghari, M., Pirsiavash, H., Brox, T.: COOT: cooperative hierarchical transformer for video-text representation learning. In: Advances in Neural Information Processing Systems (2020)
18. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: vision and language representation learning with momentum distillation. In: Advances in Neural Information Processing Systems (2021)
19. Xu, T., Feng, Z., Wu, X.J., Kittler, J.: Toward robust visual object tracking with independent target-agnostic detection and effective siamese cross-task interaction. IEEE Transactions on Image Processing, pp. 1541–1554 (2023)
20. Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4584–4593 (2016)
21. Zhang, J., Peng, Y.: Object-aware aggregation with bidirectional temporal graph for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8327–8336 (2019)
22. Ye, H., Li, G., Qi, Y., Wang, S., Huang, Q., Yang, M.H.: Hierarchical modular network for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17939–17948 (2022)
23. Xu, T., Zhu, X.-F., Wu, X.-J.: Learning spatio-temporal discriminative model for affine subspace based visual object tracking. Visual Intell. **1**(1) (2023). https://doi.org/10.1007/s44267-023-00002-1
24. Wang, B., Ma, L., Zhang, W., Liu, W.: Reconstruction network for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7622–7631 (2018)
25. Luo, H., et al.: CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
26. Xu, T., Feng, Z., Wu, X.-J., Kittler, J.: Adaptive channel selection for robust visual object tracking with discriminative correlation filters. Int. J. Comput. Vision **129**(5), 1359–1375 (2021). https://doi.org/10.1007/s11263-021-01435-1