Sports Video Analysis on Large-Scale Data

Dekun Wu[⋆1], He Zhao[⋆2], Xingce Bao[3], and Richard P. Wildes[2]
[1]University of Pittsburgh, [2]York University
[3]École Polytechnique Fédérale de Lausanne (EPFL)
dew104@pitt.edu, {zhufl, wildes}@cse.yorku.ca,
xingce.bao@alumni.epfl.ch

Abstract. This paper investigates the modeling of automated machine description on sports video, which has seen much progress recently. Nevertheless, state-of-the-art approaches fall quite short of capturing how human experts analyze sports scenes. There are several major reasons: (1) The used dataset is collected from non-official providers, which naturally creates a gap between models trained on those datasets and real-world applications; (2) previously proposed methods require extensive annotation efforts (i.e., player and ball segmentation at pixel level) on localizing useful visual features to yield acceptable results; (3) very few public datasets are available. In this paper, we propose a novel large-scale NBA dataset for Sports Video Analysis (NSVA) with a focus on captioning, to address the above challenges. We also design a unified approach to process raw videos into a stack of meaningful features with minimum labelling efforts, showing that cross modeling on such features using a transformer architecture leads to strong performance. In addition, we demonstrate the broad application of NSVA by addressing two additional tasks, namely fine-grained sports action recognition and salient player identification. Code and dataset are available at https://github.com/jackwu502/NSVA.

1 Introduction

Recently, there have been many attempts aimed at empowering machines to describe the content presented in a given video [21,12,57,40]. The particular challenge of generating a text from a given video is termed "video captioning" [2]. Sports video captioning is one of the most intriguing video captioning sub-domains, as sports videos usually contain multiple events depicting the interactions between players and objects, e.g., ball, hoop and net. Over recent years, many efforts have addressed the challenge of sports video captioning for soccer, basketball and volleyball games [40,57,54].

Despite the recent progress seen in sports video captioning, previous efforts share three major limitations. (1) They all require laborious human annotation

⋆ Equal contribution
Corresponding author: dew104@pitt.edu

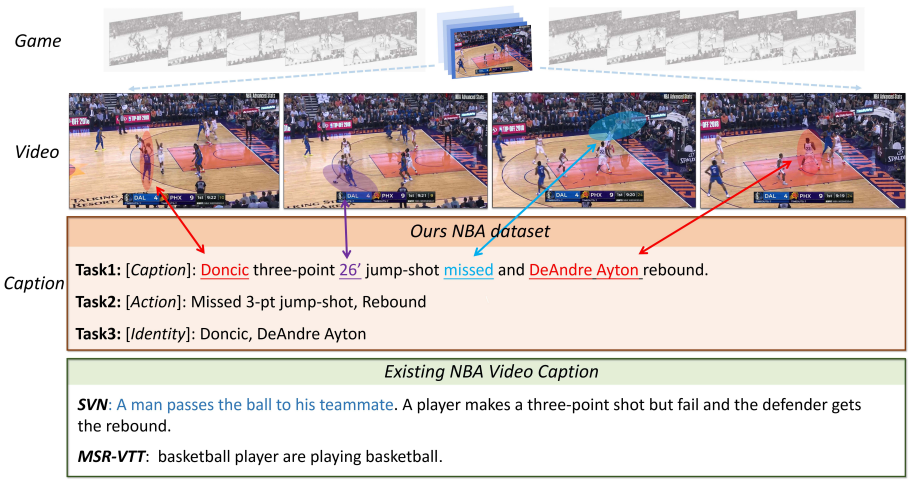


Fig. 1: Which one is more descriptive for the above professional sport game clip? Conceptual comparison between NSVA (top box) and extant basketball (NBA) video captioning datasets [57,53] (bottom box). The sentence in blue text describes a passing action, which might not be practically valuable and is not a focus of NSVA. Instead, captions in NSVA target compact information that could enable statistics counting and game analysis. Moreover, both alternative captioning approaches lack in important detail (e.g., player identities and locations).

efforts that limit the scale of data [40,57,54]. (2) Some previous efforts do not release data [40,57,54], and thereby prevent others from accessing useful data resources. (3) The collected human annotations typically lack the diversity of natural language and related intricacies. Instead, they tend to focus on details that are not interesting to human viewers, e.g. passing or dribbling activities (see Figure 1), while lacking important information (e.g. identity of performing players). In this regard, a large-scale sports video dataset that is readily accessible to researchers and annotated by professional sport analysts is very much needed. In response we propose NBA dataset for Sports Video Analysis (NSVA).

Figure 1 shows captions depicting the same sports scene from NSVA, MSR-VTT [53] and another fine-grained sports video captioning dataset, SVN [57]. Our caption is compact, focuses on key actions (e.g., *made shot*, *miss shot* and *rebound*) and is identity aware. Consequently, it could be further translated to a box score for keeping player and team statistics. SVN includes more less important actions, e.g., *passing*, *dribbling* or *standing*, which are excessively common but of questionable necessity. They neither cover player names nor essential details, e.g., shooting from 26 feet away. This characteristic of NSVA poses a great challenge as it requires models to ignore spatiotemporally dominant, yet unimportant, events and instead focus on key events that are of interest to viewers, even though they might have unremarkable visual presence. Additionally, NSVA also requires the model to identify the players whose actions will be

recorded in the box score. This characteristic adds another difficulty to NSVA and distinguishes us from all previous work, where player identification is underemphasized by only referring to "a man", "some player", "offender", etc.

Contributions. The contributions of this paper are threefold. (1) We propose a new identity-aware NBA dataset for sports video analysis (NSVA), which is built on web data, to fill the vacancy left by previous work whose datasets are neither identity aware nor publicly available. (2) Multiple novel features are devised, especially for modeling captioning as supported by NSVA, and are used for input to a unified transformer framework. Our designed features can be had with minimal annotation expense and provide complementary kinds of information for sports video analysis. Extensive experiments have been conducted to demonstrate that our overall approach is effective. (3) In addition to video captioning, NSVA is used to study salient player identification and hierarchical action recognition. We believe this is a meaningful extension to the fine-grained action understanding domain and can help researchers gain more knowledge by investigating their sports analysis models for these new aspects.

2 Related work

Video captioning aims at generating single or multiple natural language sentences based on the information stored in video clips. Researchers usually tackle this visual data-to-text problem with encoder-decoder frameworks [39,36,1,44]. Recent efforts have found object-level visual cues particularly useful for caption generation on regular videos [36,59,60,62] as well as sports videos [54,40]. Our work follows this idea to make use of detected finer visual features together with global information for professional sports video captioning.

Transformers and attention first achieved great success in the natural language domain [48,14], and then received much attention in vision research. One of the most influential pioneering works is the vision transformer (ViT) [15], which views an image as a sequence of patches on which a transformer is applied. Shortly thereafter, many tasks have found improvements using transformers, e.g., object detection [8], semantic segmentation [63,46] and video understanding [58,28,47,4]. Our work is motivated by these advances and uses transformers as building blocks for both feature extraction and video caption generation.

Sports video captioning is one of several video captioning tasks that emphasizes generation of fine-grained text descriptions for sport events, e.g., chess, football, basketball and volleyball games [11,53,18,40,57,54]. One of the biggest limitations in this area is the lack of public benchmarks. Unfortunately, none of the released video captioning datasets have a focus on sport domains. The most similar efforts to ours have not made their datasets publicly available [40,57,54], which inspires us to take advantage of webly available data to produce a new benchmark and thereby enable more exploration on this valuable topic.

Identity aware video captioning is one of the video captioning tasks that requires recognizing person identities [30,31,38]. We adopt this setting in NSVA

because successfully identifying players in a livestream game is crucial for sports video understanding and potential application to automatic score keeping. Unfortunately, the extant sports video captioning work failed to take player identities into consideration when creating their datasets. Earlier efforts that targeted player identification in professional sport scenes only experimented in highly controlled (i.e., unrealistic) environments, e.g., two teams and ten players, and has not consider incorporating identities in captioning [30,31].

Action recognition automates identification of actions in videos. Recent work has mostly focused on two sub-divisions: coarse and fine-grained recognition. The coarse level tackles basic action taxonomy and many challenging datasets are available, e.g., UCF101 [45], Kinetics [20] and ActivityNet [7]. In contrast, fine-grained distinguishes sub-classes of basic actions, with representative datasets including Diving48 [24], FineGym [42], Breakfast [22] and Epic-Kitchens [13]. Feature representation has advanced rapidly within the deep-learning paradigm (for review, see [65]) from primarily convolutional (e.g., [51,9,52,26,16]) to attention-based (e.g., [4,28]). Our study contributes to action understanding by providing a large-scale fine-grained basketball dataset that has three semantic levels as well as a novel attention-based recognition approach.

3 Data collection

Unlike previous work, we make fuller use of data that is available on the internet. We have written a webscaper to scrape NBA play-by-play data from the official website [35], which contains high resolution (e.g., 720P) video clips along with descriptions, each of which is a single event occurred in a game. We choose 132 games played by 10 teams in NBA season 2018-2019, the last season unaffected by COVID and when teams still could play with full capacity audiences, for data collection. We have collected 44,649 video clips, each of which has its associated play-by-play information, e.g., description, action and player names. We find that on the NBA website some different play-by-play information share the same video clip because there are multiple events taking place one-by-one within a short period time and the NBA just simply uses the same video clip for every event occurring in it. To avoid conflicting information in model training, the play-by-play text information sharing the same video clip is combined. We also remove the play-by-play text information that is beyond the scope of a single video clip, e.g., the points a player has scored so far in this game. This entire process is fully automated, so that we can access NBA webly data and associate video clips with captions, actions and players. Overall, our dataset consists of 32,019 video clips for fine-grained video captioning, action recognition and player identification. Additional details on dataset curation are provided in the supplement.

3.1 Dataset statistics

Table 1 shows the statistics of NSVA and two other fine-grained sports video captioning datasets. NSVA has the most sentences out of three datasets and five

Datasets	Domain	#Videos	#Sentences	#Hours	Avg. words	Accessibility	Scalability	Multi-task
SVN [54]	basketball	5,903	9,623	7.7	8.8	✗	✗	✗
SVCDV [40]	volleyball	4,803	44,436	36.7	-	✗	✗	✗
NSVA	basketball	**32,019**	**44,649**	**84.8**	6.5	✓	✓	✓

Table 1: The statistics of NSVA and comparison to other fine-grained sports video captioning datasets.

times more videos than both SVN and SVCDV. The biggest strength of NSVA is its public accessibility and scalability. Both SVN and SVCDV datasets are neither publicly available nor scalable because heavy maunal annotation effort is required in their creation. In contrast, NSVA is built on data that already existed on the internet; so, everyone who is interested can directly download and use the data by following our guidelines. Indeed, the 132 games that we chose to use only accounts for 10.7% of total games in NBA season 2018-2019. There is more data being produced everyday as NBA teams keep playing and sharing their data. Note that some other datasets also contain basketball videos, e.g., MSR-VTT [53] and ActivityNet [7]. However, they only provide coarse-level captions (see example in Figure 1) and include very limited numbers of videos, e.g, ActivityNet has 74 videos for basketball and they are all from amateur play, not professional.

Table 2 shows the data split of NSVA. We hold 32 games out from 132 games to form validation set and test set, each of which contains 16 games. All clips and texts belonging to a single game are assigned to the same data split. When choosing what data split a game is assigned to, we ensure that every team match-up has been seen at least once in the training set. For example, Phoenix Suns play four games against San Antonio Spurs in NBA season 2018-2019. We put two games in the training set, one in the validation set and one in the test set.

NSVA also supports two additional vision tasks, namely fine-grained action recognition and key player identification. We adopt the same data curation strategy as captioning and show the number of distinct action or player name categories in the rightmost two columns of Table 2. When being compared with other find-grained sport action recognition datasets, e.g., Diving48 (48 categories) and Finegym (530 categories), ours is in the middle place (172 categories) in terms of number of actions and is the largest regarding the basketball sub-domain.

Videos				Sentences				Games				Teams	Actions	Identities
train	val	test	total	train	val	test	total	train	val	test	total	all-sets	all-sets	all-sets
24k	3.9k	3.9k	32k	33.6k	5.5k	5.5k	44.6k	100	16	16	132	10	172	184

Table 2: Data split detail of our dataset.

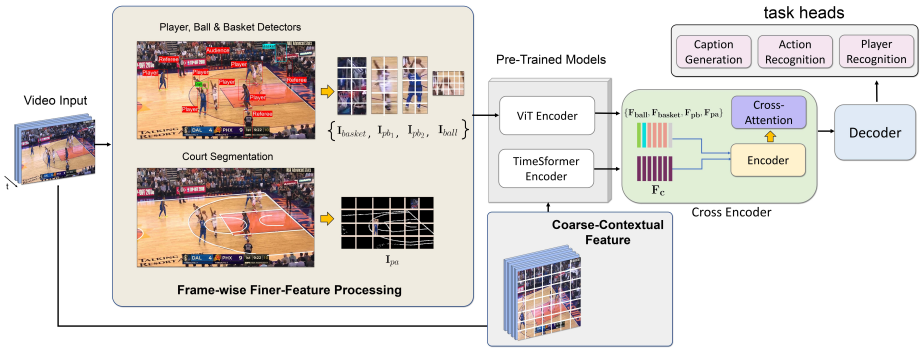


Fig. 2: Pipeline of our proposed approach for versatile sports video understanding. First, raw video clips (left) are processed into two types of finer visual information, namely object detection (including ball, players and basket), and court-line segmentation, all of which are cropped, grided and channelled into a pre-trained vision transformer model for feature extraction. Second, these heterogeneous features are aggregated and cross-encoded with the global contextual video representation extracted from TimeSformer (middle). Third, a transformer decoder is used with task-specific heads to recursively yield results, be it as video captions, action recognition or player identification (right).

4 Architecture design

Problem formulation. We seek to predict the correct sequence of word captions as one-hot vectors, $\{\mathbf{y}\}$, whose length is arbitrary, given the observed input clip $X \in \mathbb{R}^{H\times W\times 3\times N}$ consisting of N RGB frames of size $H \times W$ sampled from the original video.

Overall structure. As our approach relies on feature representations extracted from multiple orthogonal perspectives, we adopt the framework of UniVL [32], a network designed for cross feature interactive modeling, as our base model. It consists of four transformer backbones that are responsible for coarse feature encoding, fine-grained feature encoding, cross attention and decoding, respectively. In the following, we step-by-step detail our multi-level feature extraction, integrated feature modeling and decoder.

4.1 Course contextual video modeling

In most video captioning efforts a 3D-CNN has been adopted as the fundamental unit for feature extraction, e.g., S3D [52,54,40]. More recent work employed a transformer architecture in tandem [32]. Inspired by TimeSformer [4], which is solely built on a transformer block and has shown strong performance on several action recognition datasets, we substitute the S3D part of UniVL with this new model as video feature extractor. Correspondingly, we decompose each frame into F non-overlapping patches, each of size $P \times P$, such that the F patches span the entire frame, i.e., $F = HW/P^2$. We flatten these patches into vectors

and channel them into several blocks comprised of linear-projection, multihead-self-attention and layer-normalization, in both spatial and temporal axes, which we shorten as

$$\mathbf{F}_c = \text{TimeSformer}\left(X\right), \tag{1}$$

where $\mathbf{F}_c \in \mathbb{R}^{N\times d}$, d is the feature dimension and X is an input clip.

Transformer blocks have less strong inductive priors compared to convolutional blocks, so they can more readily model long-range spatiotemporal information with their self-attention mechanism in a large-scale data learning setting. We demonstrate the strong performance of TimeSformer features in Sec. 5.

4.2 Fine-grained objects of interest modeling

One limitation of solely using TimeSformer features is that we might lose important visual details, e.g., ball, players and basket, after resizing 1280×720 images to 224×224, the size that TimeSformer encoder needs. Such loss can be important because NSVA requires modeling main players' identities and their actions to generate an accurate caption. To remedy this issue, we use an object detector to capture objects of interest that contain rich regional semantic information complementary to the global semantic feature provided by TimeSformer. We extract 1,000 image frames from videos in the training set and annotate bounding boxes for basket and ball and fine-tune on the YOLOv5 model [19] to have a joint ball-basket object detector. This pre-trained model returns ball and basket crops from original images, i.e., $\mathbf{I}_{ball}$ and $\mathbf{I}_{basket}$.

For player detector, we simply use the YOLOv5 model trained on the MS-COCO dataset [27] to retrieve a stack of player crops, $\{\mathbf{I}_{player}\}$. As our caption is identity-aware, we assume that players who have touched the ball during a single play are more likely to be mentioned in captions. Thus, we only keep the detected players that have overlap with a detected ball, e.g., each player crop, $\mathbf{I}_{player}$, is given a confidence score, C, of 1 otherwise 0; in particular, if $\text{IoU}\left(\mathbf{I}_{player_i}, \mathbf{I}_{ball}\right) > 0 : C = 1;$ else $: C = 0$. Player crops that have $C = 1$ will be selected for later use, $\mathbf{I}_{pb}$. Even though the initially detected players, $\{\mathbf{I}_{player}\}$, potentially are contaminated by non–players (e.g., referees, audience members), our ball-focused confidence scores tend to filter out these distractors.

After getting bounding boxes of ball, players intersecting with the ball and basket, we crop these objects from images and feed them to a vision transformer, ViT [15], for feature extraction,

$$\mathbf{f}_{ball} = \text{ViT}\left(\mathbf{I}_{ball}\right),\ \mathbf{f}_{basket} = \text{ViT}\left(\mathbf{I}_{basket}\right),\ \mathbf{f}_{pb} = \text{ViT}\left(\mathbf{I}_{pb}\right), \tag{2}$$

where $\mathbf{f}_{ball}$, $\mathbf{f}_{pb}$ and $\mathbf{f}_{basket}$ are features of d dimension extracted from cropped ball image, $\mathbf{I}_{ball}$, player with ball image, $\mathbf{I}_{pb}$, and basket image, $\mathbf{I}_{basket}$, respectively. We re-group features from every second in the correct time order to have $\mathbf{F}_{ball}$, $\mathbf{F}_{basket}$ and $\mathbf{F}_{pb}$, which all are of dimensions $\mathbb{R}^{m\times d}$.

Discussion. Compared with previous work that either require pixel-level annotation in each frame to segment each player, ball and background [54], or

person-level annotation that needs professional sport knowledge to recognize each player's action such as setting, spiking and blocking [40], our annotation scheme is very lightweight. The annotation only took two annotators less than five hours to draw bounding boxes for ball and basket in 1,000 selected image frames from the training set. Compared to the annotation procedure that requires months of work for experts with extensive basketball knowledge [54], our approach provides a more affordable, replicable and scalable option. Note that these annotations are only for training the detectors; the generation of the dataset per se is completely automated; see Sec. 3.

4.3 Position-aware module

NSVA supports modeling estimation of the distance from where the main player's actions take place to the basket. As examples, "Lonnie Walker missed **2'** cutting layup shot" and "Canaan **26'** 3PT Pullup Jump Shot", where the numbers in bold denote the distance between the player and basket. Notably, distance is strongly correlated with action; e.g., players cannot make a 3PT shot at two-foot distance from the basket. While estimating such distances is important for action recognition and caption generation, it is non-trivial owing to the need to estimate separation between two 3D objects from their 2D image projections.

Instead of explicitly making such prediction directly on raw video frames, we take advantage of prior knowledge that basketball courtlines are indicators of object's location. We use a pix2pix network [17] trained on synthetic data [64] to generate courtline segmentation given images. We overlay the detected player with ball and basket region, while blacking out other areas. Figure 2 shows an exemplar image, $\mathbf{I}_{pa}$, after such processing. We feed these processed images to ViT for feature extraction, i.e., $\mathbf{F_{pa}} = \text{ ViT } (\mathbf{I}_{pa})$, where $\mathbf{F}_{pa} \in \mathbb{R}^{m \times d}$ are ViT features extracted from position-aware image $\mathbf{I}_{pa}$.

4.4 Visual transformer encoder

After harvesting the video, ball, basket and courtline features, we are ready to feed them into the coarse encoder as well as the finer encoder for self-attention. This step is necessary as the used backbones (i.e., ViT and TimeSformer) only perform attention on frames within one second; there is no communication between different timestamps. For this purpose, we use one transformer to encode video feature, $\mathbf{F}_c \in \mathbb{R}^{N \times d}$ (1), and another transformer to encode aggregated finer features, $\mathbf{F}_f \in \mathbb{R}^{M \times 2d}$, which is from the concatenation of position-aware feature, $\mathbf{F}_{pa}$, and the summation of object-level features. Empirically, we find summation sufficient, i.e.,

$$\mathbf{F}_f = \text{CONCAT}(\text{SUM}(\mathbf{F}_{ball}, \mathbf{F}_{basket}, \mathbf{F}_{pb}), \mathbf{F}_{pa}) \tag{3}$$

The overall encoding process is given as

$$\mathbf{V}_c = \text{Transformer}\left(\mathbf{F}_c\right), \mathbf{V}_f = \text{Transformer}\left(\mathbf{F}_f\right), \tag{4}$$

where $\mathbf{V}_c \in \mathbb{R}^{n \times d}$ and $\mathbf{V}_f \in \mathbb{R}^{m \times d}$.

4.5 Cross encoder for feature fusion

The coarse and fine encoders mainly focus on separate information. To make them fully interact, we follow existing work and adopt a cross encoder [32], which takes coarse features, $\mathbf{V}_c$, and fine features, $\mathbf{V}_f$, as input. Specifically, these features are combined along the sequence dimension via concatenation and a transformer is used to generate the joint representation, i.e.,

$$\mathbf{M} = \text{Transformer}(\text{CONCAT}(\mathbf{V_c}, \mathbf{V_f})), \tag{5}$$

where $\mathbf{M}$ is the final output of the encoder. To generate a caption, a transformer decoder is used to attend $\mathbf{M}$ and output text autoregressively, cf., [47,6,41].

4.6 Learning and inference

Finally, we calculate the loss as the sum of negative log likelihood of correct caption at each step according to

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log P_\theta \left(y_t \mid y_{<t}, \mathbf{M}\right), \tag{6}$$

where θ is the trainable parameters, $y_{<t}$ is the ground-truth words sequence before step t and y_t is the ground truth word at step t.

During inference, the decoder autoregressively operates a beam search algorithm [33] to produce results, with beam size set empirically; see Sec. 5.1.

4.7 Adaption to other tasks

In NSVA, action and identity also are sequential data. So, we adopt the same model, shown in Figure 2, for all three tasks and swap the caption supervision signal in (6), $y_{1:t}$, with either one-hot action labels or player name labels. Similarly, inference operates beam search decoding. Details are in the supplement.

5 Empirical evaluation

5.1 Implementation details

We use hidden state dimension of 768 for all encoders/decoders. We use the BERT [14] vocabulary augmented with 356 action types and player names entries. The transformer encoder, cross-attention and decoder are pretrained on a large instructional video dataset, Howto100M [34]. We keep the pre-trained model and fine tune it on NSVA, as we found the pre-trained weights speed up model convergence. The maximum number of frames for the encoder and the maximum output length are set to 30. The number of layers in the feature encoder, cross encoder and decoder are 6, 3 and 3, respectively. We use the Adam

Model	Feature	C	M	B@1	B@2	B@3	B@4	R_L
MP-LSTM [50]	S3D	0.500	0.153	0.325	0.236	0.167	0.121	0.332
TA [55]	S3D	0.546	0.156	0.331	0.242	0.175	0.128	0.340
Transformer [43]	S3D	0.572	0.161	0.346	0.254	0.181	0.131	0.357
UniVL* [32]	S3D	0.717	0.192	0.441	0.309	0.226	0.169	0.401
Our Model	T	0.956	0.217	0.467	0.363	0.274	0.209	0.468
	S3D+BAL+BAS+PB+PA	0.986	0.227	0.479	0.371	0.281	0.216	0.466
	T+BAL	0.931	0.228	0.496	0.383	0.289	0.220	0.484
	T+BAS	1.023	0.232	0.500	0.387	0.292	0.223	0.486
	T+PB	1.055	0.231	0.500	0.387	0.292	0.223	0.487
	T+PA	1.064	0.238	0.511	0.398	0.301	0.231	0.498
	T+BAL+BAS	1.074	0.243	0.508	0.398	0.306	0.237	0.499
	T+BAL+BAS+PB	1.096	0.242	0.519	0.408	0.312	0.242	0.506
	T+BAS+BAL+PB+PA	**1.139**	**0.243**	**0.522**	**0.410**	**0.314**	**0.243**	**0.508**

Table 3: Performance comparison of our model vs. alternative video captioning models on the NSVA test set. T denotes TimeSformer feature. BAL, BAS and PB denote ViT features for ball, basket and player with ball, respectively. PA is the position-aware feature. *As our model adopts the framework of UniVL as backbone, results in the row of UniVL+S3D equals to those of our model only using S3D features.

optimizer with an initial learning rate of 3e-5 and employ a linear decay learning rate schedule with a warm-up strategy. We used a batch size of 32 and trained our model on a single Nvidia Tesla T4 GPU for 12 epochs over 6 hours. The hyperparameters were chosen based on the top performer on the validation set.

In testing we adopt beam search [33] with beam size 5. For extraction of the TimeSformer feature, we sample video frames at 8 fps. For extraction of other features, we sample at 12 vs 4 fps when the ball is vs is not detected in the basket area. We record the time when the ball first is detected and keep 100 frames before and after. This step saves about 70% storage space compared to sampling the entire video at 8 fps, but still keeps the most important frames.

5.2 Video captioning

Baseline and evaluation metrics. The main task of NSVA is video captioning. To assess our proposed approach, we compare our results with four state-of-the-art video captioning systems: MP-LSTM [50], TA [55], Transformer [43] and UniVL [32] on four widely-used evaluation metrics: CIDEr (C) [49], Bleu (B) [37], Meteor (M) [3] and Rouge-L (R_L) [25]. Results are shown in Table 3. To demonstrate the effectiveness of our approach against the alternatives, we train these models on NSVA using existing codebases [56,32].

Main results. Comparing results in the first two rows with results in other rows of Table 3, we see that transformer models outperform LSTM models, which confirms the superior capability of a transformer on the video captioning task. Moveover, it is seen that TimeSformer features achieve much better results

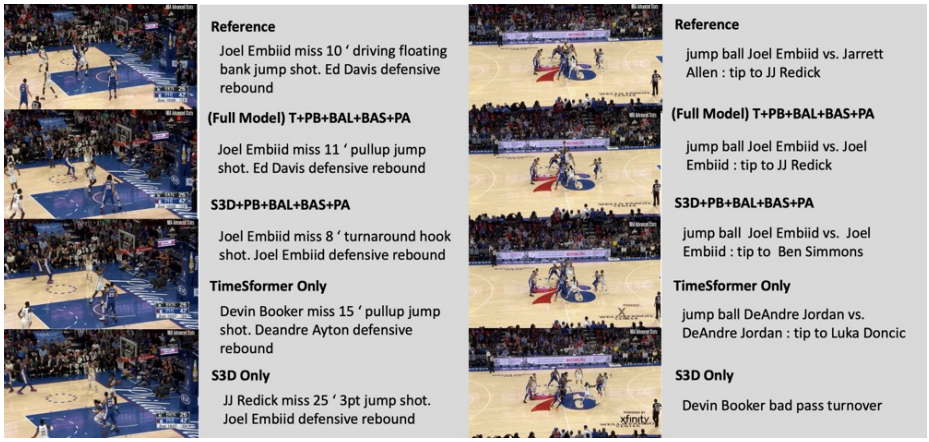


Fig. 3: Qualitative analysis of captions generated by our proposed approach and others. It is seen that captions from our full approach are the most close to references.

compared to S3D in modeling video context. We conjecture that this is due to its ability to model long spatiotemporal dependency in videos; see 4^{th} and 5^{th} rows. This result suggests that TimeSformer features are not only useful for video understanding tasks but also video captioning. Comparing results on the 4^{th} and 6^{th} rows, we find that after fusing S3D features with those extracted by our proposed modules (but not the TimeSformer), improvements are seen on all metrics. A possible explanation is that our features add additional semantic information (e.g., pertaining to ball, player and court) and thereby lead to higher quality text. The best result is achieved by fusing TimeSformer features with our proposed features. These results suggest that (1) TimeSformer features are well suited to video captioning and (2) our proposed features can be fused with a variety of features for video understanding to improve performance further. From the 7^{th} to final row of Table 3, we ablate our finer-grained features. It is seen that our model benefits from every proposed finer module, and when combining all modules, we observe the best result; see last row. This documents the effectiveness of our proposed method for the video captioning task on NSVA. More discussions on the empirical results can be found in the supplement.

Qualitative analysis. Figure 3 shows two example outputs generated by four different models, as compared to the ground-truth reference. From the left example output, we see that our full model is able to generate a high quality caption, albeit with relatively minor mistakes. After replacing the TimeSformer features with S3D features, the model fails to identify the player who gets the rebound and mistakes a jump shot for a hook shot. When using TimeSformer or S3D feature alone, the result further deteriorate by misidentifying all players. We also notice that our devised features, i.e., PB+BAL+BAS+PA, can greatly help capture a player's position, e.g., with 10' as the reference, models

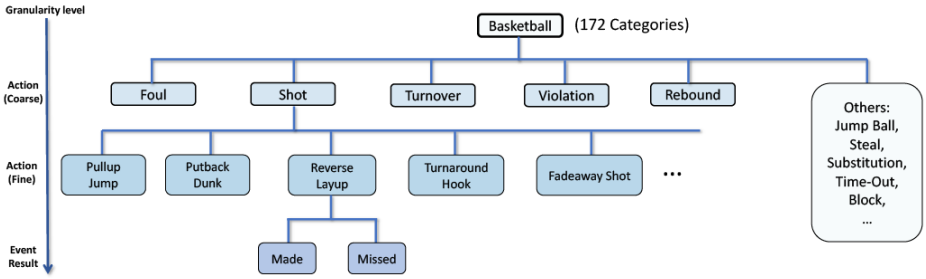


Fig. 4: Visualization of a sub-tree from our fine-grained basketball action space. There are 172 fine-grained categories that comprise three levels of sport event details: **Action-C** (coarse), **Action-F** (fine) and **Action-E** (event). Some categories have finer descendants (e.g., *Shot*), while others are solitary (e.g., *Jump Ball* and *Block*). The full list of action categories is in the supplement.

with PB+BAL+BAS+PA features output 11' and 8', compared to 15' and 25' output by TimeSformer only and S3D only.

The right column shows an example where all models successfully recognize the action, i.e., jump shot, except the S3D only model. Our full model can identify most players but still mistakes Jarret Allen for Joel Embiid. As we will discuss in Sec. 5.4, player identification is the bottleneck of our model as it is trained with a very weak supervision signal, which points to future research.

5.3 Fine-grained basketball action recognition

As elaborated in Sec. 3, NSVA has massive video clips that cover almost every moment of interest, and these events have been provided by the NBA for the purpose of statistics tracking, which allows fine-grained action recognition. A glimpse of how our action labels are hierarchically organized is shown at Figure 4.

Action hierarchy. NSVA enjoys three levels of granularity in the basketball action domain. (1) On the coarsest level, there exist 14 actions that describe the on-going sport events from a very basic perspective. Some representative examples include: { *Shot, Foul, Turnover* }. (2) If further dividing the coarse actions into their finer sub-divisions, we can curate 124 fine-grained actions. Taking the shot category as an example, it has the following sub-categories: { *Shot Dunk*, *Shot Pullup Jumpshot*, *Shot Reverse Layup*, etc. }. All of these finer actions enrich the coarse ones with informative details (e.g., diverse styles for the same basketball movement). (3) On the finest level, there exists 24 categories that depicts the overall action from the event perspective, which includes the coarse action name, the fine action style and the overall event result, e.g., { *Shot-Pullup-Jumpshot-Missed* }. Thanks to the structured labelling, NSVA can support video action understanding on multiple granularity levels. We demonstrate some preliminary results using our proposed approach in Table 4.

					Action-C			Action-F			Action-E		
Feature-backbone	PB	BAL	BAS	PA	SR↑	Acc.↑	mIoU↑	SR↑	Acc.↑	mIoU↑	SR↑	Acc.↑	mIoU↑
TimeSformer	✓	✓	✓	✓	**60.14**	**61.20**	**66.61**	**46.88**	**51.25**	57.08	**37.67**	**42.34**	**46.45**
TimeSformer	✓	✓	✓	-	60.02	60.79	65.33	46.42	50.64	**57.19**	36.44	42.29	42.14
TimeSformer	✓	✓	-	-	58.06	60.31	63.71	44.31	49.01	55.78	34.53	39.34	46.45
TimeSformer	✓	-	-	-	57.74	58.13	60.48	44.20	50.18	55.91	34.50	39.14	42.72
TimeSformer	-	-	-	-	55.83	58.01	60.19	42.55	49.66	53.81	33.63	37.50	40.84
S3D	-	-	-	-	54.46	57.91	59.91	41.92	48.81	53.77	33.09	37.11	40.77

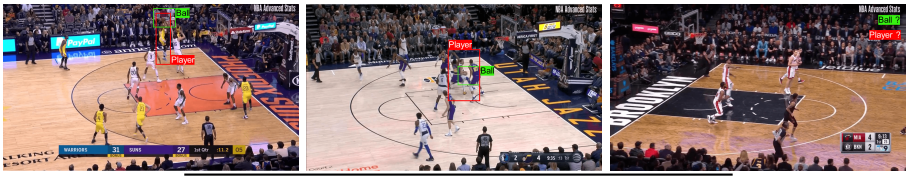
Table 4: Action recognition accuracy (%) on NSVA at all granularities.

Evaluation. As exemplified in Figure 1, our action labels do not always assign a single ground-truth label to a clip. In fact, they contain as many actions as happens within the length of a unit clip. The example in Figure 1 shows a video clip that has two consecutive actions, i.e., [*3-pt Jump-Shot Missed* → *Defensive Rebound*]. To properly evaluate our results in this light, we adopt metrics from efforts studying instructional videos [10,5,61], and report: (1) mean Intersection over Union (mIoU), (2) mean Accuracy (Acc.) and (3) Success Rate (SR). Detailed explanation can be found in the supplement. We provide action recognition results using the same feature design introduced in Sec. 4 and provide an ablation study on the used features.

Results on multiple granularity recognition. From the results in Table 4, we can summarize several observations: (1) Overall, actions in NSVA are quite challenging to recognize, as the best result on the coarsest level only achieves 61.2% accuracy (see columns under Action-C). (2) When the action space is further divided into sub-actions, the performance becomes even weaker (e.g., 51.25% for Action-F and 42.43 % for Action-E), meaning that subtle and challenging differences can lead to large drops in recognizing our actions. (3) TimeSformer features perform better than S3D counterparts at all granularity levels, which suggests NSVA benefits from long-term modeling. (4) We observe solid improvements by gradually incorporating our devised finer features, which once again demonstrates the utility of our proposed approach.

5.4 Player identification

We adopt the same training and evaluation strategy as in action recognition to measure the performance of our model on player identification, due to these tasks having the same format, i.e., a sequence of player names involved in the depicted action; Fig. 5 has results. Resembling observations in the previous subsection, we find the quality of identified player names increases as we add more features and our full approach (top row) once again is the best performer. It also is seen that the results on all metrics are much worse than those of action recognition, cf., Table 4. To explore this discrepancy, we study some failure cases in the images along the top of Fig. 5. It is seen that failure can be mostly attributed to blur, occlusion from unrelated regions and otherwise missing decisive information.



Feature-backbone	PB	BAL	BAS	PA	SR ↑	Acc ↑	mIoU ↑
TimeSformer	✓	✓	✓	✓	**4.63**	**6.97**	6.86
TimeSformer	✓	✓	✓	-	4.20	6.83	**6.89**
TimeSformer	✓	✓	-	-	4.17	6.45	6.68
TimeSformer	✓	-	-	-	3.97	6.33	6.52
TimeSformer	-	-	-	-	3.66	5.98	6.07
S3D	-	-	-	-	3.57	5.91	5.49

Fig. 5: (Top) Visual explanations revealing difficulty in player identification. Left: Although our detector captures the ball and player correctly, the face, jersey and size of the key player are barely recognizable due to blur. Middle: The detected player area is crowded and the ball handler is occluded by defenders. Right: A case where the ball is missing; thus, the model cannot find decisive information on the key player. (Bottom) Player identification results in percentage (%) with our full approach and ablations on choice of features.

6 Conclusion

In this work, we create a large-scale sports video dataset (NSVA) supporting multiple tasks: video captioning, action recognition and player identification. We propose a unified model to tackle all tasks and outperform the state of the art by a large margin on the video captioning task. The creation of NSVA only relies on webly data and needs no extra annotation. We believe NSVA can fill the opening for a benchmark in fine-grained sports video captioning, and potentially stimulate the application of automatic score keeping.

The bottleneck of our model is player identification, which we deem the most challenging task in NSVA. To this end, a better algorithm is needed, e.g., opportunistic player recognition when visibility allows, with subsequent tracking for fuller inference of basketball activities. There also are two additional directions we will explore: (1) We will investigate more advanced video feature representations (e.g., Video Swin transformer [29]) on NSVA and compare to TimeSformer. (2) Prefix Multi-task learning [23] has been proposed to learn several tasks in one model. Ideally, a model can benefit from learning to solve all tasks and gain extra performance boost on each task. We will investigate NSVA in the Prefix Multi-task learning setting with our task head.

Acknowlegement. The authors thank Professor Adriana Kovashka for meaningful discussions in the early stages and Professor Hui Jiang for proofreading and valuable feedback. This research was supported in part by a NSERC grant to Richard P. Wildes and a CFREF VISTA Graduate Scholarship to He Zhao.

References

1. 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 CVPR (2019)
2. Aafaq, N., Mian, A., Liu, W., Gilani, S.Z., Shah, M.: Video description: A survey of methods, datasets, and evaluation metrics. ACM Comput. Surv. **52**(6) (Oct 2019)
3. Banerjee, S., Lavie, A.: METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of ACL (2005)
4. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding. In: Proceedings of ICML (2021)
5. Bi, J., Luo, J., Xu, C.: Procedure planning in instructional videos via contextual modeling and model-based policy learning. In: Proceedings of ICCV (2021)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020)
7. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: ActivityNet: A large-scale video benchmark for human activity understanding. In: Proceedings of CVPR (2015)
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of ECCV (2020)
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the Kinetics dataset. In: proceedings of CVPR (2017)
10. Chang, C.Y., Huang, D.A., Xu, D., Adeli, E., Fei-Fei, L., Niebles, J.C.: Procedure planning in instructional videos. In: Proceedings of ECCV (2020)
11. Chen, D., Dolan, W.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of ACL (2011)
12. Chen, S., Song, Y., Zhao, Y., Qiu, J., Jin, Q., Hauptmann, A.G.: RUC+CMU: system report for dense captioning events in videos. CoRR **abs/1806.08854** (2018)
13. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The Epic-Kitchens dataset. In: Proceedings of ECCV (2018)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL (2019)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
16. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of ICCV (2019)
17. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of CVPR (2017)
18. Jhamtani, H., Gangal, V., Hovy, E., Neubig, G., Berg-Kirkpatrick, T.: Learning to generate move-by-move commentary for chess games from large-scale social forum data. In: Proceedings of ACL (2018)
19. Jocher, G., Stoken, A., Chaurasia, A., Borovec, J., Xie, T., Kwon, Y., Michael, K., Changyu, L., Fang, J., Skalski, P., Hogan, A., Nadar, J., Mammana, L., Wang, A., Fati, C., Montes, D., Hajek, J., Diaconu, L., Minh, M.T., Haoyang, W.: Yolov5:v6.0 (2021). https://doi.org/10.5281/zenodo.5563715

20. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The Kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
21. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: Proceedings ICCV (2017)
22. Kuehne, H., Arslan, A., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Proceedings of CVPR (2014)
23. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of ACL (2021)
24. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: Proceedings of ECCV (2018)
25. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
26. Lin, J., Gan, C., Han, S.: TSM: Temporal shift module for efficient video understanding. In: Proceedings of ICCV (2019)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proceedings of ECCV (2014)
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of ICCV (2021)
29. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of CVPR (2022)
30. Lu, W.L., Ting, J.A., Little, J.J., Murphy, K.P.: Learning to track and identify players from broadcast sports videos. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(7), 1704–1716 (2013)
31. Lu, W.L., Ting, J.A., Murphy, K.P., Little, J.J.: Identifying players in broadcast sports videos using conditional random fields. In: proceedings of CVPR (2011)
32. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Chen, X., Zhou, M.: UniVL: A unified video and language pre-training model for multimodal understanding and generation. CoRR **abs/2002.06353** (2020)
33. Medress, M.F., Cooper, F.S., Forgie, J.W., Green, C., Klatt, D.H., O'Malley, M.H., Neuburg, E.P., Newell, A., Reddy, D., Ritea, B., et al.: Speech understanding systems: Report of a steering committee. Artificial Intelligence **9**(3), 307–316 (1977)
34. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of CVPR (2019)
35. NBA: Official website, `https://www.nba.com`
36. Pan, B., Cai, H., Huang, D.A., Lee, K.H., Gaidon, A., Adeli, E., Niebles, J.C.: Spatio-temporal graph for video captioning with knowledge distillation. In: Proceedings of CVPR (2020)
37. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of ACL (2002)
38. Park, J.S., Darrell, T., Rohrbach, A.: Identity-aware multi-sentence video description. In: Proceedings of ECCV (2020)
39. Pei, W., Zhang, J., Wang, X., Ke, L., Shen, X., Tai, Y.W.: Memory-attended recurrent network for video captioning. In: Proceedings of CVPR (2019)
40. Qi, M., Wang, Y., Li, A., Luo, J.: Sports video captioning via attentive motion representation and group relationship modeling. IEEE Transactions on Circuits and Systems for Video Technology **30**(8), 2617–2633 (2019)

41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of ICML (2021)
42. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of CVPR (2020)
43. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of ACL (2018)
44. Shi, B., Ji, L., Niu, Z., Duan, N., Zhou, M., Chen, X.: Learning semantic concepts and temporal alignment for narrated video procedural captioning. In: Proceedings of MM (2020)
45. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
46. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of ICCV (2021)
47. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: VideoBert: A joint model for video and language representation learning. In: Proceedings of ICCV (2019)
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
49. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDER: Consensus-based image description evaluation. In: Proceedings of CVPR (2015)
50. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. In: Proceedings of NAACL (2015)
51. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.V.: Temporal segment networks: Towards good practices for deep action recognition. In: Proceedings of ECCV (2016)
52. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of ECCV (2018)
53. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: Proceedings of CVPR (2016)
54. Yan, Y., Zhuang, N., Ni, B., Zhang, J., Xu, M., Zhang, Q., Zhang, Z., Cheng, S., Tian, Q., Xu, Y., Yang, X., Zhang, W.: Fine-grained video captioning via graph-based multi-granularity interaction learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(2), 666–683 (2022)
55. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: Proceedings of ICCV (2015)
56. Yehao, L., Yingwei, P., Jingwen, C., Ting, Y., Tao, M.: X-modaler: A versatile and high-performance codebase for cross-modal analytics. In: Proceedings of MM (2021)
57. Yu, H., Cheng, S., Ni, B., Wang, M., Zhang, J., Yang, X.: Fine-grained video captioning for sports narrative. In: Proceedings of CVPR (2018)
58. Zhang, C., Gupta, A., Zisserman, A.: Temporal query networks for fine-grained video understanding. In: Proceedings of CVPR (2021)
59. Zhang, J., Peng, Y.: Object-aware aggregation with bidirectional temporal graph for video captioning. In: Proceedings of CVPR (2019)
60. Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., Zha, Z.J.: Object relational graph with teacher-recommended learning for video captioning. In: Proceedings of CVPR (2020)

61. Zhao, H., Hadji, I., Dvornik, N., Derpanis, K.G., Wildes, R.P., Jepson, A.D.: P3iv: Probabilistic procedure planning from instructional videos with weak supervision. In: Proceedings of CVPR (2022)
62. Zheng, Q., Wang, C., Tao, D.: Syntax-aware action targeting for video captioning. In: Proceedings of CVPR (2020)
63. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of CVPR (2021)
64. Zhu, L., Rematas, K., Curless, B., Seitz, S.M., Kemelmacher-Shlizerman, I.: Reconstructing NBA players. In: Proceedings of ECCV (2020)
65. Zhu, Y., Li, X., Liu, C., Zolfaghari, M., Xiong, Y., Wu, C., Zhang, Z., Tighe, J., Manmatha, R., Li, M.: A comprehensive study of deep video action recognition. arXiv preprint arXiv:2012.06567 (2020)