Identity-Aware Multi-sentence Video Description

Jae Sung Park[1(✉)], Trevor Darrell[2], and Anna Rohrbach[2]

[1] Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, USA
jspark96@cs.washington.edu
[2] University of California, Berkeley, Berkeley, USA

Abstract. Standard video and movie description tasks abstract away from person identities, thus failing to link identities across sentences. We propose a multi-sentence Identity-Aware Video Description task, which overcomes this limitation and requires to re-identify persons locally within a set of consecutive clips. We introduce an auxiliary task of Fill-in the Identity, that aims to predict persons' IDs consistently within a set of clips, when the video descriptions are given. Our proposed approach to this task leverages a Transformer architecture allowing for coherent joint prediction of multiple IDs. One of the key components is a gender-aware textual representation as well an additional gender prediction objective in the main model. This auxiliary task allows us to propose a two-stage approach to Identity-Aware Video Description. We first generate multi-sentence video descriptions, and then apply our Fill-in the Identity model to establish links between the predicted person entities. To be able to tackle both tasks, we augment the Large Scale Movie Description Challenge (LSMDC) benchmark with new annotations suited for our problem statement. Experiments show that our proposed Fill-in the Identity model is superior to several baselines and recent works, and allows us to generate descriptions with locally re-identified people.

1 Introduction

Understanding and describing videos that contain multiple events often requires establishing "who is who": who are the participants in these events and who is doing what. Most of the prior work on automatic video description focuses on individual short clips and ignores the aspect of participants' identity. In particular, prior works on movie description tend to replace all character identities with a generic label SOMEONE [35,41]. While reasonable for individual short clips, it becomes an issue for longer video sequences. As shown in Fig. 1, descriptions that contain SOMEONE would not be satisfying to visually impaired users, as they do not unambiguously convey who is engaged in which action in video.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58589-1_22) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12366, pp. 360–378, 2020.
https://doi.org/10.1007/978-3-030-58589-1_22

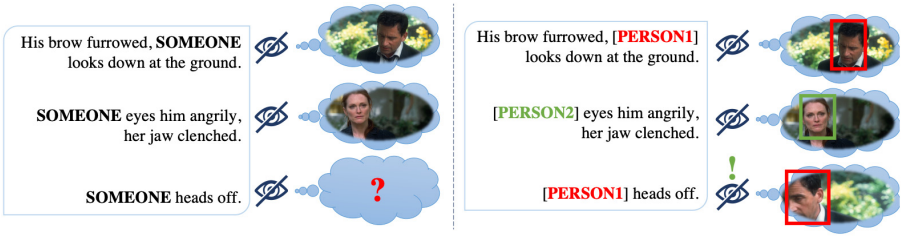

Fig. 1. Compare video description with SOMEONE labels vs. **Identity-Aware Video Description**: in the first case it may be difficult for a visually impaired person to follow what is going on in the video, while in the second case it becomes clear who is performing which action.

Several prior works attempt to perform person re-identification [5,36,39] in movies and TV shows, sometimes relying on associated subtitles or textual descriptions [27,31,40]. Most such works take the "linking tracks to names" problem statement, i.e. trying to name all the detected tracks with proper character names. Others like [29] aim to "fill in" the character proper names in the given ground-truth movie descriptions.

In this work, we propose a different problem statement, which does not require prior knowledge of movie characters and their appearance. Specifically, we group several consecutive movie clips into sets and aim to establish person identities *locally* within each set of clips. We then propose the following two tasks. First, given ground-truth descriptions of a set of clips, the goal is to fill in person identities in a coherent manner, i.e. to predict the same ID for the same person within a set (see Fig. 2). Second, given a set of clips, the goal is to generate video descriptions that contain corresponding local person IDs (see Fig. 1). We refer to these two tasks as **Fill-in the Identity** and **Identity-Aware Video Description**. The first (auxiliary) task is by itself of interest, as it requires to establish person identities in a multi-modal context of video and description.

We experiment with the Large Scale Movie Description Challenge (LSMDC) dataset and associated annotations [34,35], as well as collect more annotations to support our new problem statement. We transform the global character information into local IDs within each set of clips, which we use for both tasks.

Fill-in the Identity. Given textual descriptions of a sequence of events, we aim to fill in the person IDs in the blanks. In order to do that, two steps are necessary. First, we need to attend to a specific person in the video by relating visual observations to the textual descriptions. Second, we need to establish links within a set of blanks by relating corresponding visual appearances and textual context. We learn to perform both steps jointly, as the only supervision available to us is that of the identities, not the corresponding visual tracks. Our key idea is to consider an entire set of blanks jointly, and exploit the mutual relations between the attended characters. We thus propose a Transformer model [42] which jointly infers the identity labels for all blanks. Moreover, to support this

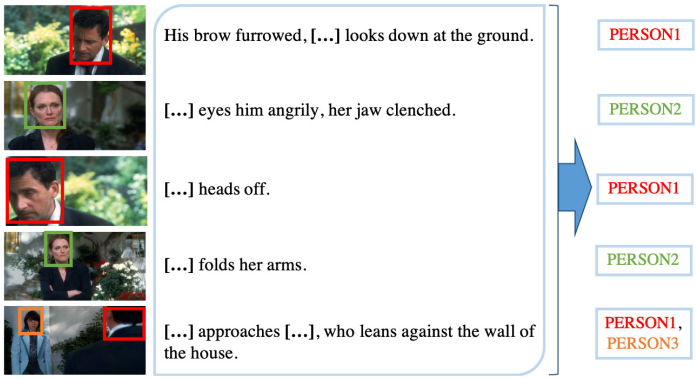

Fig. 2. Example of the **Fill-in the Identity** task.

process we make use of one additional cue available to us: the gender of the person in question. We train a text-based gender classifier, which we integrate in our model, along with an additional visual gender prediction objective, which aims to recognize gender based on the attended visual appearance.

Identity-Aware Video Description. Given a set of clips, we aim to generate descriptions with local IDs. Here, we take a two-stage approach, where we first obtain the descriptions with SOMEONE labels as a first step, and next apply our Fill-in the Identity method to give SOMEONEs their IDs. We believe there is a potential in exploring other models that would incorporate the knowledge of identities into generation process, and leave this to future work.

Our contributions are as follows. (1) We introduce a new task of **Identity-Aware Video Description**, which extends prior work in that it aims to obtain multi-sentence video descriptions with local person IDs. (2) We also introduce a task of **Fill-in the Identity**, which, we hope, will inform future directions of combining identity information with video description. (3) We propose a Transformer model for this task, which learns to attend to a person and use gender evidence along with other visual and textual cues to correctly fill in the person's ID. (4) We obtain state-of-art results in the **Fill-in the Identity** task, compared to several baselines and two recent methods. (5) We further leverage this model to address **Identity-Aware Video Description** via a two-stage pipeline, and show that it is robust enough to perform well on the generated descriptions.

2 Related Work

Video Description. Automatic video description has attracted a lot of interest in the last few years, especially since the arrival of deep learning techniques [2,9,22,23,33,44,54,58]. Here are a few trends present in recent works [20,30,47,62]. Some works formulate video description as a reinforcement learning problem [19,28,49]. Several methods address grounding semantic concepts during description

generation [52,59,61]. A few recent works put an emphasis on the use of syntactic information [14,45]. New datasets have also been proposed [12,50], including a work that focuses on dense video captioning [17], where the goal is to temporally localize and caption individual events.

In this work, we generate multi-sentence descriptions for long video sequences, as in [32,37,55]. This is different from dense video captioning, where one does not need to obtain one coherent multi-sentence description. Recent works that tackle multi-sentence video description include [56], who generate fine-grained sport narratives, [53], who jointly localize events and decide when to generate the following sentence, and [25], who introduce a new inference method that relies on multiple discriminators to measure the quality of multi-sentence descriptions.

Person Re-identification. Person re-identification aims to recognize whether two images of a person are the same individual. This is a long standing problem in computer vision, with numerous deep learning based approaches introduced over the years [5,26,36,39,46]. We rely on [36] as our face track representation.

Connections to Prior Work. Next, we detail how our work compares to the most related prior works.

Identity-Aware Video Description: Closely related to ours is the work of [34]. They address video description of *individual* movie clips with grounded and co-referenced (re-identified) people. In their problem statement re-identification is performed w.r.t. *a single previous clip* during description generation. Unlike [34], we address *multi-sentence* video description, which requires consistently re-identifying people over *multiple clips at once* (on average 5 clips).

Fill-in the Identity: Our task of predicting *local* character IDs for a set of clips given ground-truth descriptions with blanks, is related to the work of [29]. However, they aim to fill in *global* IDs (proper names). In order to learn the global IDs, they use 80% of each movie for training. Our problem statement is different, as it requires no access to the movie characters' appearance during training: we maintain disjoint training, validation and test movies. A number of prior works attempt to link all the detected face tracks to global character IDs in TV shows and movies [3,11,15,21,27,31,38,40], which is different from our problem statement that tries to fill character IDs locally with textual guidance. We compare to two recent approaches to Fill-in the Identity in Sect. 5.2.

3 Connecting Identities to Video Descriptions

An integral part of understanding a story depicted in a video is to establish who are the key participants and what actions they perform over the course of time. Being able to correctly link the repeating appearances of a certain person could potentially help follow the story line of this person. We first address the task of **Fill-in the Identity**, where we aim to solve a related problem: fill in the persons' IDs based on the given video descriptions with blanks.

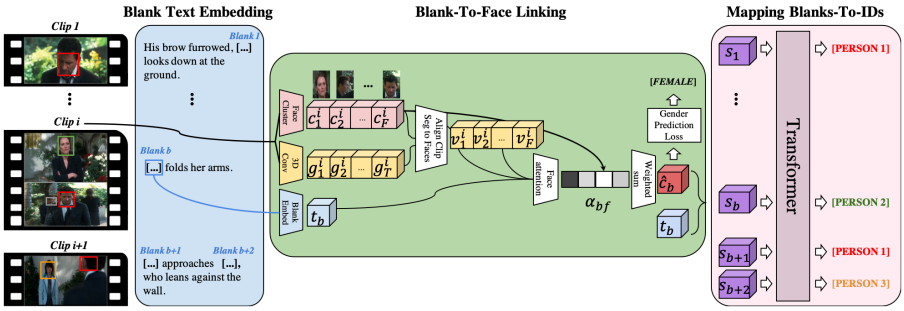

Fig. 3. Overview of our approach to **Fill-in the Identity** task. See Sect. 3.1.

Our approach is centered around two key ideas: joint prediction of IDs via a Transformer architecture [42], supported by gender information inferred from textual and visual cues (see Fig. 3). We then present our second task, **Identity-Aware Video Description**, which aims to generate multi-sentence video descriptions with local person IDs. We present a two-stage pipeline, where our baseline model gives us multi-sentence descriptions with SOMEONE labels, after which we leverage the **Fill-in the Identity** auxiliary task to link the predicted SOMEONE entities.

3.1 Fill-in the Identity

For a set of video clips V_i and their descriptions with blanks $D_i, i = 1, 2..., N$, we aim to fill in character identities that are locally consistent within the set. We first detect faces that appear in each clip and cluster them based on their visual appearance. Then, for every blank in a sentence, we attend over the face cluster centers using visual and textual context, to find the cluster best associated with the blank. We process all the blanks sequentially and pass their visual and textual representations to a Transformer model [42], which analyzes the entire set at once and predicts the most probable sequence of character IDs (Fig. 3).

Visual Representation. Now, we describe the details of getting local face descriptors and other global visual features for a clip V_i. We detect all the faces in every frame of the clip, using the face detector by [60]. Then, we extract 512-dim face feature vectors with the FaceNet model [36][1] trained on the VGGFace2 dataset [5]. The feature vectors are clustered using DBSCAN algorithm [10], which does not require specifying a number of clusters as a parameter. We take the mean of face features in each cluster, resulting in F face feature vectors $(c_1^i, ..., c_F^i)$.

In addition to the face features, we extract spatio-temporal features that describe the clip semantically. These features help the model locate where to look for the relevant face cluster for a given blank. We extract I3D [6] features and apply mean pooling across a fixed number of T segments following [48],

[1] https://github.com/davidsandberg/facenet.

giving us a sequence $(g_1^i, ..., g_T^i)$. We then associate each face cluster with the best temporally aligned segment as follows. For each face cluster c_f^i, we keep track of its frame indices and get a "center" index. Dividing this index by the total number of frames in a clip gives us a relative temporal position of the cluster r_f^i, $0 <= r_f^i < 1$. We get the corresponding segment index $t_f^i = \lfloor r_f^i * T \rfloor$ and obtain the global visual context $v_f^i = g_{t_f^i}^i$. We concatenate face cluster features c_f^i with the associated global visual context v_f^i as our final visual representation.

Filling in the Blanks. Suppose there are B blanks in the set of N sentences D_i[2]. One way to fill in these blanks is to train a language model, such as BERT [8], by masking the blanks to directly predict the character IDs. As we aim to incorporate visual information in our model, we take the following approach.

First, each blank b in a sentence D_i has to receive a designated textual encoding. We use with a pretrained BERT model, which has been shown effective for numerous NLP tasks. Instead of relying on a generic pretrained model, we train it to predict the *gender* corresponding to each blank, which often can be inferred from text. For example, in Fig. 2, one can infer that the person in the first clip is male due to the phrase "His brow". We process all sentences in the set jointly. To get a representation for each blank, we access output embedding from the [CLS] token, a special sentence classification token in [8] whose representation captures the meaning of the entire sentence, over all sentences, and a hidden state of the last layer associated with the specific blank token. Note, that the same [CLS] token is used for all blanks in the set. The final representation t_b is a concatenation of the [CLS] and the blank token representation.

For each clip V_i, we obtain F face cluster representations $(c_1^i, ...c_F^i)$, which we combine with the corresponding clip level representations $(v_1^i, ...v_F^i)$. To find the best matching face cluster for the blank b, we predict the attention weights α_{bf} over all clusters in the clip based on the t_b, and compute a weighted sum over the face clusters, $\hat{c_b}$:

$$\begin{aligned} e_{bf} &= W_{\alpha_2} \tanh(W_{\alpha_1}[c_f^i; v_f^i; t_b]), \\ \alpha_{bf} &= \frac{\exp(e_{bf})}{\sum_{k=1}^{F} \exp(e_{bk})}, \hat{c_b} = \sum_{f=1}^{F} \alpha_{bf} c_f \end{aligned} \tag{1}$$

We concatenate the visual representation $\hat{c_b}$ with t_b as the final representation for the blank: $s_b = [\hat{c_b}; t_b]$.

Given a set of B blanks represented by $(s_1, ..., s_B)$ we now aim to link the corresponding identities to each other. Instead of making pairwise decisions w.r.t. matching and non-matching blanks, we want to predict an entire sequence of IDs jointly. We thus choose the Transformer [42] architecture to let the self-attention mechanisms model multiple pairwise relationships at once. Specifically, we train a Transformer with $(s_1, ..., s_B)$ as inputs and local person IDs $(l_1, ..., l_B)$ as outputs. As we fill in the blanks in a sequential manner, we prevent the future blanks from

[2] Some sentences may have multiple blanks, others may have none.

impacting the previous blanks by introducing a causal masking in the encoder. We train the entire model end-to-end and learn the attention mechanism in Eq. 1 jointly with the ID prediction. Denoting θ as all the parameters in the model, the loss function we minimize is:

$$L_{character}(\theta) = -\sum_{b=1}^{B} p_\theta(l_b \mid s_1, ..., s_{b-1}, l_1, ..., l_{b-1}) \tag{2}$$

We explore the effect of an additional component, a *gender prediction loss* L_{gender}, that forces the attended visual representation to be gender-predictive. We add a single layer perceptron that takes the predicted feature $\hat{c_b}$ and aims to recognize the gender g_b for the blank b. The final loss function we minimize is as follows:

$$\begin{aligned} L_{gender}(\theta) &= -\sum_{b=1}^{B} p_\theta(g_b \mid \hat{c_b}) \\ L(\theta) &= L_{character} + \lambda_{gen} L_{gender} \end{aligned} \tag{3}$$

where $\hat{c_b}$ is calculated in Eq. 1 and λ_{gen} is a hyperparameter.

We also notice that it is possible to boost the performance of our Transformer model by a simple training data augmentation. Note that there are various ways to split the training data into clip sequences with length N: one can consider a non-overlapping segmentation, i.e. $\{1, ...N\}, \{N+1, ...2N\}, ...$ or additionally add all the overlapping sets $\{2, ...N+1\}, \{3, ...N+2\}, ...$ Since we predict the local IDs, every such set would result in a unique data point, meaning using all the possible sets can increase the amount of training data by a factor of N.

3.2 Identity-Aware Video Description

Here, given a set of N clips V_i, we aim to predict their descriptions D_i that would also contain the relevant local person IDs. First, we follow prior works [12,25] to build a multi-sentence description model with SOMEONE labels. It is an LSTM based decoder that takes as input a visual representation and a sentence generated for a previous clip. Here, our visual representation for V_i is I3D [6] and Resnet-152 [13] mean pooled temporal segments. Once we have obtained a multi-sentence video description with SOMEONE labels, we process the generated sentences with our Fill-in the Identity model. We demonstrate that this approach is effective, although the Fill-in the Identity model is only trained on ground-truth descriptions. Note, that this two-stage pipeline could be applied to any video description method.

4 Dataset

As our main test-bed, we choose the Large Scale Movie Description Challenge (LSMDC) [35], while leveraging and extending the character annotations

Table 1. Statistics for our tasks, based on the LSMDC dataset. See Section 4.

	Movies	Sentences	Sets	Blanks
Training	153	101,079	20,283	87,604
Validation	12	7,408	1,486	6,457
Public Test	17	10,053	2,018	8,431

from [34]. They have marked every sentence in the MPII Movie Description (MPII-MD) dataset where a person's proper name (e.g. *Jane*) is mentioned, and labeled the person specific pronouns *he, she* with the associated names (e.g. *he* is *John*). For each one out of 94 MPII-MD movies, we are given a list of all unique identities and their genders. We extend these annotations to 92 additional movies, covering the entire LSMDC dataset (except for the Blind Test set).

We use these annotations as follows. (1) We drop the pronouns and focus on the underlying IDs. (2) We split each movie into sets of consecutive 5 clips (the last set in a movie may contain less than 5 clips). (3) We relabel global IDs into local IDs *within each set*. E.g. if we encounter a sequence of IDs *Jane, John, Jane, Bill*, it will become *PERSON1, PERSON2, PERSON1, PERSON3*. This relabeling is applied for both tasks that we introduce in this work.

We provide dataset statistics, including number of movies, individual sentences, sets and blanks in Table 1[3]. Around 52% of all blanks correspond to PERSON1, 31% – to PERSON2, 12% – to PERSON3, 4% – to PERSON4, 1% or less – to PERSON5, PERSON6, etc. This shows that the clip sets tend to focus on a single person, but there is still a challenge in distinguishing PERSON1 from PERSON2, PERSON3, ... (up to PERSON11 in training movies).

In our experiments we use the LSMDC Validation set for development, and LSMDC Public Test set for final evaluation.

5 Experiments

5.1 Implementation Details

For each clip, we extract I3D [6] pre-trained on the Kinetics [16] dataset and Resnet-152 [13] features pre-trained on the ImageNet [7] dataset. We mean pool them temporally to get $T = 5$ segments [48]. We detect on average 79 faces per clip (and up to 200). In the DBSCAN algorithm used to cluster faces, we set $\epsilon = 0.2$, which is the maximum distance between two samples in a cluster. Clustering the faces within each clip results in about 2.2 clusters per clip, and clustering over a set results in 4.2 clusters per set. BERT [8] models use the BERT-base model architecture with default settings as in [51]. Transformer [42]

[3] Note, that the reported number of training clip sets reflects the default non-overlapping "segmentation", as done for validation and test movies. One is free to define the training clip sets as arbitrary sets of 5 consecutive clips.

has a feedforward dimension of 2048 and 6 self-attention layers. We train the Fill-in the Identity model for 40 epochs with learning rate 5e−5 with hyperparameter $\lambda_{gender} = 0.2$. We train the baseline video description model for 50 epochs with learning rate 5e−4. We fix batch size as 16 across all experiments, where each batch contains a set of clips and descriptions.

5.2 Fill-in the Identity

Evaluation Metrics. First, we discuss the metrics used to evaluate the **Fill-in the Identity** task. Given a sequence of blanks and corresponding ground-truth IDs, we consider all unique pairwise comparisons between the IDs. A pair is labeled as "Same ID" if the two IDs are the same, and "Different ID" otherwise. We obtain such labeling for the ground-truth and predicted IDs. Then, we can compute a ratio of the matching labels between the ground-truth and predicted pairs, e.g. if 6 out of 10 pair labels match, the prediction gets an accuracy 0.6[4]. The final accuracy is averaged across all sets. When define like this, we obtain an *instance-level* accuracy over ID pairs ("Inst-Acc"). Note, that it is important to correctly predict both "Same ID" and "Different ID" labels, which can be seen as a 2-class prediction problem. The instance-level accuracy does not distinguish between these two cases. Thus, we introduce a *class-level* accuracy, where we separately compute accuracy over the two subsets of ID pairs ("Same-Acc", "Diff-Acc") and report the harmonic mean between the two ("Class-Acc").

Baselines and Ablations. Table 2 summarizes our experiments on the LSMDC Validation set. We include two simple baselines: "The same ID" (all IDs are the same) and "All different IDs" (all IDs are distinct: 1, 2, ...). "GT Gender as ID" directly uses *ground truth* male/female gender as a character ID (Person 1/2), and serves as an upper-bound for gender prediction. We consider two vision-only baselines, where we cluster all the detected faces within a set of 5 clips, and pick a random cluster ("Random Face Cluster") or the most frequent cluster ("Most Frequent Face Cluster") within a clip for each blank. We also consider a language-only baseline "BERT Character LM", which uses a pretrained BERT model to directly fill in all the blanks. Then we include our Transformer model with our BERT Gender pretrained model. We show the effect of our training data augmentation and use augmentation in the following versions. Finally, we study the impact of adding each visual component ("+ Face." and "+ Video"), and introduce our vision-based gender loss (full model).

We make the following observations. (1) Instance accuracy for all the same/all distinct IDs provides the insight into how the pairs are distributed (40.7% of all pairs belong to "Same ID" class, 59.3.7% – to "Different ID" class). Neither is a good solution, getting Class-Acc 0. (2) Our Transformer model with BERT Gender representation improves over the vanilla BERT Character model (57.9

[4] This resembled pairwise precision/recall used in clustering [1]. However, these are not applicable in our scenario as they can not handle singleton clusters (with one element). Thus, we compute pairwise accuracy instead.

Table 2. Fill-in the Identity accuracy of several baselines, our full method and its ablations on the LSMDC Validation set. We report the predicted ID accuracy at class and instance level, as well as gender accuracy. See Section 5.2 for details.

Method	Same Acc	Diff Acc	Class Acc	Inst Acc	Gen Acc
The same ID	100.0	0.0	0.0	40.7	-
All different IDs	0.0	100.0	0.0	59.3	-
GT Gender as ID	100.0	43.0	60.1	65.5	-
V: Random Face Cluster	54.8	52.0	53.4	55.3	-
V: Most Frequent Face Cluster	48.3	68.8	56.7	61.6	-
L: BERT Character LM	57.9	65.1	57.9	66.4	-
L: Transf. + BERT Gen. LM	57.3	68.9	62.6	67.9	80.3
L: Transf. + BERT Gen. LM + Aug.	60.7	68.7	64.4	69.6	81.8
L+V: Transf. + BERT Gen. LM + Aug. + Face.	62.8	68.0	65.3	69.7	81.8
L+V: Transf. + BERT Gen. LM + Aug. + Face + Video.	64.8	66.6	65.7	69.6	81.8
L+V: Transf. + BERT Gen. LM + Aug. + Face + Video + Gen. Loss	63.5	68.4	65.9	69.8	83.0

vs. 62.6 in Class-Acc). (3) This is also higher than 60.1 of "GT Gender as ID", i.e. our model relies on other language cues besides gender. (4) Training with our data augmentation scheme further improves Class-Acc to 64.4. (5) Introducing face features boosts the Class-Acc to 65.3, and adding video features improves it to 65.7. (7) Finally, visual gender prediction loss leads to the overall Class-Acc of 65.9. Note, that the instance-level accuracy (Inst-Acc) does not always reflect the improvements as it may favor the majority class ("Different ID").

We also report gender accuracy (Gen Acc) for the variants of our model (last 4 rows in Table 2). For models without the visual gender loss, we report the accuracy based on their BERT language model trained for gender classification. We see that data augmentation on the language side helps improve gender accuracy (80.3 vs 81.8). Incorporating visual representation with the gender loss boosts the accuracy further (81.8 vs 83.0).

Human Performance. We also assess human performance in this task in two scenarios: with and without seeing the video. The former provides an overall upper-bound accuracy, while the latter gives an upper-bound for the text-only models. We randomly select 200 sets of Test clips and ask 3 different Amazon Mechanical Turk (AMT) workers to assign the IDs to the blanks. For each set we compute a *median* accuracy across the 3 workers to remove the outliers and report the average accuracy over all sets. Table 3 reports the obtained human performance and the corresponding accuracy of our model on the same 200 sets. Human performance gets a significant boost when the workers can see the video, indicating

Table 3. Fill-in the Identity human performance and our method evaluated on 200 random Test clips sets.

Method	Same Acc	Diff Acc	Class Acc	Inst Acc
Human w/o video	56.5	78.5	65.7	70.0
Human	83.9	90.0	86.8	87.0
Ours	64.6	70.7	67.5	70.3

that *video provides many valuable cues*, and not everything can be inferred from text. Our full method outperforms "Human w/o video" but falls behind the final "Human" performance.

Comparison to State-of-the-Art. Since the data for our new tasks has been made public, other research groups have reported results on it, including the works by Yu et al. [57] and Brown et al. [4][5]. Yu et al. propose an ensemble over two models: Text-Only and Text-Video. The Text-Only model builds a pairwise matrix for a set of blanks and learns to score each pair by optimizing a binary cross entropy loss. The Text-Video model considers two tasks: linking blanks to tracks and linking tracks to tracks. The two tasks are trained separately, with the help of additional supervision from an external dataset [29], using triplet margin loss. While Yu et al. use gender loss to pre-train their language model, we introduce gender loss on the visual side.

Brown et al. train a Siamese network on positive (same IDs) and negative (different IDs) pairs of blanks. The network relies on an attention mechanism over the face features to identify the relevant person given the blank encoding.

Table 4. Fill-in the Identity accuracy of our method and two recent works on the LSMDC Test set.

Method	Same Acc	Diff Acc	Class Acc	Inst Acc
Yu et al. [57]	26.4	87.3	40.6	65.9
Brown et al. [4]	33.6	81.0	47.5	64.8
Ours text only	56.0	71.2	62.7	68.0
Ours	60.6	70.0	64.9	69.6

Table 4 reports the results on the LSMDC Test set. As we see, our approach significantly outperforms both methods. To gain further insights, we analyze some differences in behavior between these methods and our approach.

We take a closer look at the distribution of the predicted IDs (PERSON1, PERSON2, ...) by our method, and the two other approaches. Figure 4 provides a histogram over the reference data and the compared approaches. We can see that our predictions align well with the true data distribution, while the two other methods struggle to capture the data distribution. Notably, both methods favor more diverse IDs (2, 3, ...), failing to link many of the re-occurring character appearances. This may be in part due to the difference in the objective used in our approach vs. the others. While they implement a binary classifier that selects the best matching face track for each blank, we use Transformer to fill in the blanks jointly, allowing us to better capture both local and global context.

Figure 5 provides a qualitative example, comparing our approach, Yu et al. and Brown et al. As suggested by our analysis, these two methods often predict diverse IDs instead of recognizing the true correspondences.

[5] Note, that we have corrected some errors, affecting about 3% of the annotations. While Yu et al. [57] and Brown et al. [4] have trained their models on the old annotations, all reported results are obtained on the *corrected* test set.

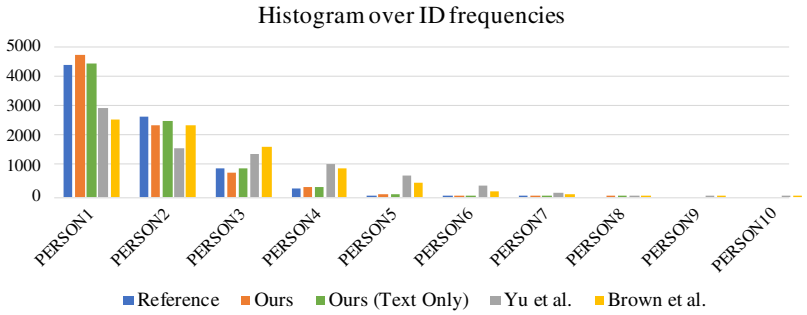

Fig. 4. Fill-in the Identity: histogram over the frequencies of predicted IDs for our method, its text-only version and two SOTA works. See Sect. 5.2.

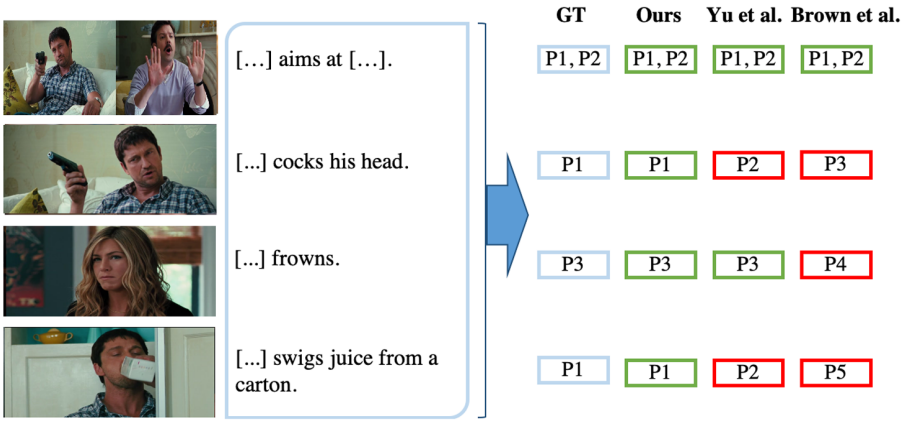

Fig. 5. Qualitative example for **Fill-in the Identity** task, comparison between our approach and two recent methods. Correct/incorrect predictions are labeled with green/red, respectively. P1, P2, ... are person IDs. See Sect. 5.2 for details. (Color figure online)

Ours vs. Ours Text Only. In Fig. 6, we compare our full model vs. its text-only version. After having seen the two characters in the first clip (P1 and P2), our full model recognizes that the man and woman appearing in a next set of clips are the same two characters, and successfully links them as P1 in the second and P2 in the third and fourth clip with the correct genders. In the last clip with two characters, the full model is also able to visually ground the same woman as "heading out" and assign the ID as P2. On the other hand, the text-only model cannot tell that the first two characters appear in the next set of clips without a visual signal, and incorrectly assigns the blanks as different character IDs, P3 and P4, after the second blank. The text only model also fails to link the character in third and last clip as the same ID due to the limited information available in textual descriptions alone. This shows that the task is hard for the

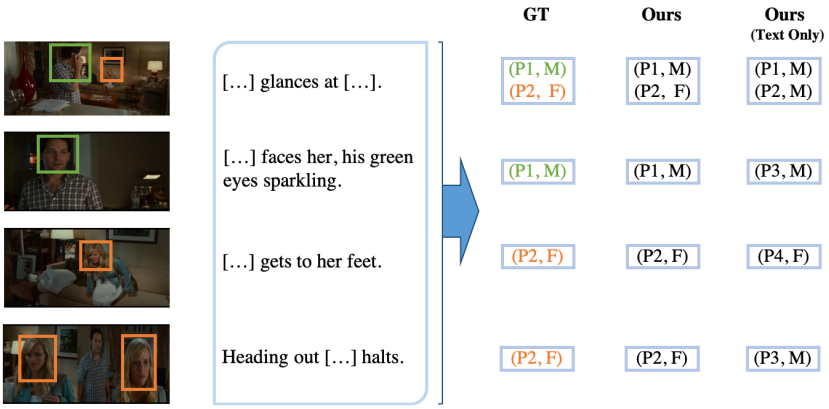

Fig. 6. Qualitative example for **Fill-in the Identity** task, comparison between our final approach with visual representation and a text-only ablation ("Transf. + BERT Gender LM + Augm." in Table 2). We include the predicted character ID (P1, P2, ...) and gender for each blank. See Sect. 5.2 for details.

Table 5. Identity-Aware Video Description scores for our method on the LSMDC Test set. See Sect. 5.3 for details.

Method	Per set, MAX score		
	METEOR	BLEU@4	CIDEr-D
Same ID	9.22	1.59	6.31
All different IDs	8.84	1.38	6.06
Ours Text-Only	10.29	1.74	6.88
Ours	10.38	1.75	6.95

text-only model, while *our final model learns to incorporate visual information successfully.*

5.3 Identity-Aware Video Description

Finally, we evaluate our two-stage approach for the **Identity-Aware Video Description** task. One issue with directly evaluating predicted descriptions with the character IDs is that the evaluation can strongly penalize predictions that do not closely follow the ground-truth. Consider an example with ground-truth *[P1] approaches [P2]. [P2] gets up.* vs. prediction *[P1] is approached by [P2]. [P1] stands up.* As we see, direct comparison would yield a low score for this prediction, due to different phrasing which leads to different local person IDs. If we instead consider an ID permutation *[P2] approaches [P1]. [P1] gets up.*,

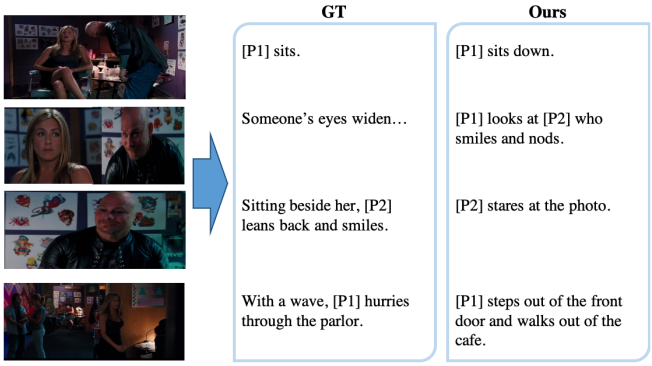

Fig. 7. Qualitative example for the **Identity-Aware Video Description** task. P1, P2, ... are person IDs. See Sect. 5.3 for details.

we can get a better match with the prediction. Thus, we consider all the possible ID permutations as references, evaluate our prediction w.r.t. all of them, and choose the reference that gives the highest BLEU@4 to compute the final scores. The caption with the best ID permutation is then used to evaluated at set level using the standard automatic scores (METEOR [18], BLEU@4 [24], CIDEr-D [43]).

Results. In Table 5, we compare results from our captioning model (Sect. 3.2) with different Fill-in the Identity approaches to fill in the IDs on the LSMDC Test set. Our model outperforms the baseline approaches, including Ours Text-Only model. This confirms that our Fill-in the Identity model successfully uses visual signal to perform well, on both ground truth and predicted sentences.

Figure 7 provides an example output of our two-stage approach. We show the ground truth descriptions with person IDs on the left, and our generated descriptions with the predicted IDs on the right. Our Fill-in the Identity model consistently links [P1] as the woman who "sits down", "looks at [P2]", and "steps out", while [P2] as the man who "smiles" and "stares" across the clips.

While we experiment with a fairly straightforward video description model, our two-stage approach can enable character IDs if applied to any model.

6 Conclusion

In this work we address the limitation of existing literature on automatic video and movie description, which typically ignores the aspect of person identities.

Our main effort in this paper is on the Fill-in the Identity task, namely filling in the local person IDs in the given descriptions of clip sequences. We propose a new approach based on the Transformer architecture, which first learns to attend to the faces that best match the blanks, and next jointly establishes links between them (infers their IDs). Our approach successfully leverages gender information, inferred both from textual and visual cues. Human performance on the Fill-in the Identity task shows the importance of visual information, as humans perform much better when they see the video. While we demonstrate that our approach benefits from visual features (higher ID accuracy and gender accuracy, better ID distribution), future work should focus on better ways of incorporating visual information in this task. Finally, we compare to two state-of-the-art multi-modal methods, showing a significant improvement over them.

We also show that our Fill-in the Identity model enables us to use a two-stage pipeline to tackle the Identity-Aware Video Description task. While this is a simple approach, it is promising to see that our model can handle automatically generated descriptions, and not only ground-truth descriptions. We hope that our new proposed tasks will lead to more research on bringing together person re-identification and video description and will encourage future works to design solutions that go beyond a two-stage pipeline.

Acknowledgements. The work of Trevor Darrell and Anna Rohrbach was in part supported by the DARPA XAI program, the Berkeley Artificial Intelligence Research (BAIR) Lab, and the Berkeley DeepDrive (BDD) Lab.

References

1. Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S., Mooney, R.J.: Model-based overlapping clustering. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 532–537 (2005)
2. Baraldi, L., Grana, C., Cucchiara, R.: Hierarchical boundary-aware neural encoder for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
3. Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Finding actors and actions in movies. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2013)
4. Brown, A., Albanie, S., Liu, Y., Nagrani, A., Zisserman, A.: LSMDC V2 challenge presentation. In: 3rd Workshop on Closing the Loop Between Vision and Language (2019)
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: a dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE (2018)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255. IEEE (2009)

8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
10. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231. ACM (1996)
11. Everingham, M., Sivic, J., Zisserman, A.: "hello! my name is... buffy" - automatic naming of characters in TV video. In: Proceedings of the British Machine Vision Conference (BMVC) (2006)
12. Gella, S., Lewis, M., Rohrbach, M.: A dataset for telling the stories of social media videos. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 968–974 (2018)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
14. Hou, J., Wu, X., Zhao, W., Luo, J., Jia, Y.: Joint syntax representation learning and visual cue translation for video captioning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8918–8927 (2019)
15. Jin, S., Su, H., Stauffer, C., Learned-Miller, E.: End-to-end face detection and cast grouping in movies using Erdos-Renyi clustering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5276–5285 (2017)
16. Kay, W., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
17. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 706–715 (2017)
18. Lavie, M.D.A.: Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), p. 376 (2014)
19. Li, L., Gong, B.: End-to-end video captioning with multitask reinforcement learning. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2019)
20. Li, Y., Yao, T., Pan, Y., Chao, H., Mei, T.: Jointly localizing and describing events for dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7492–7500 (2018)
21. Miech, A., Alayrac, J.B., Bojanowski, P., Laptev, I., Sivic, J.: Learning from video and text via large-scale discriminative clustering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5257–5266 (2017)
22. Pan, P., Xu, Z., Yang, Y., Wu, F., Zhuang, Y.: Hierarchical recurrent neural encoder for video representation with application to captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
23. Pan, Y., Yao, T., Li, H., Mei, T.: Video captioning with transferred semantic attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
24. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2002)

25. Park, J.S., Rohrbach, M., Darrell, T., Rohrbach, A.: Adversarial inference for multi-sentence video description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
26. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference (BMVC) (2015)
27. Parkhi, O.M., Rahtu, E., Zisserman, A.: It's in the bag: stronger supervision for automated face labelling. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops) (2015)
28. Pasunuru, R., Bansal, M.: Reinforced video captioning with entailment rewards. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2017)
29. Pini, S., Cornia, M., Bolelli, F., Baraldi, L., Cucchiara, R.: M-VAD names: a dataset for video captioning with naming. Multimedia Tools Appl. **78**(10), 14007–14027 (2019)
30. Rahman, T., Xu, B., Sigal, L.: Watch, listen and tell: multi-modal weakly supervised dense event captioning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8908–8917 (2019)
31. Ramanathan, V., Joulin, A., Liang, P., Fei-Fei, L.: Linking people in videos with "their" names using coreference resolution. In: Proceedings of the European Conference on Computer Vision (ECCV) (2014)
32. Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., Schiele, B.: Coherent multi-sentence video description with variable level of detail. In: Proceedings of the German Conference on Pattern Recognition (GCPR) (2014)
33. Rohrbach, A., Rohrbach, M., Schiele, B.: The long-short story of movie description. In: Proceedings of the German Conference on Pattern Recognition (GCPR) (2015)
34. Rohrbach, A., Rohrbach, M., Tang, S., Oh, S.J., Schiele, B.: Generating descriptions with grounded and co-referenced people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
35. Rohrbach, A., et al.: Movie description. Int. J. Comput. Vis. (IJCV) **123**(1), 94–120 (2017)
36. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
37. Shin, A., Ohnishi, K., Harada, T.: Beyond caption to narrative: video captioning with multiple sentences. In: Proceedings of the IEEE IEEE International Conference on Image Processing (ICIP) (2016)
38. Sivic, J., Everingham, M., Zisserman, A.: "who are you?"-learning person specific classifiers from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
39. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
40. Tapaswi, M., Baeuml, M., Stiefelhagen, R.: "knock! knock! who is it?" probabilistic person identification in TV-series. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
41. Torabi, A., Pal, C., Larochelle, H., Courville, A.: Using descriptive video services to create a large data source for video annotation research. arXiv:1503.01070 (2015)
42. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (NIPS) (2017)

43. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
44. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence - video to text. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
45. Wang, B., Ma, L., Zhang, W., Jiang, W., Wang, J., Liu, W.: Controllable video captioning with POS sequence guidance based on gated fusion network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2641–2650 (2019)
46. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: NormFace: L2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM International Conference on Multimedia. ACM (2017)
47. Wang, J., Jiang, W., Ma, L., Liu, W., Xu, Y.: Bidirectional attentive fusion with context gating for dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7190–7198 (2018)
48. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 20–36. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_2
49. Wang, X., Chen, W., Wu, J., Wang, Y.F., Wang, W.Y.: Video captioning via hierarchical reinforcement learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4213–4222 (2018)
50. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: a large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
51. Wolf, T., et al.: Huggingface's transformers: state-of-the-art natural language processing. arXiv abs/1910.03771 (2019)
52. Wu, X., Li, G., Cao, Q., Ji, Q., Lin, L.: Interpretable video captioning via trajectory structured localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
53. Xiong, Y., Dai, B., Lin, D.: Move forward and tell: a progressive generator of video descriptions. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
54. Yao, L., et al.: Describing videos by exploiting temporal structure. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
55. 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 (CVPR) (2016)
56. Yu, H., Cheng, S., Ni, B., Wang, M., Zhang, J., Yang, X.: Fine-grained video captioning for sports narrative. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6006–6015 (2018)
57. Yu, Y., Chung, J., Kim, J., Yun, H., Kim, G.: LSMDC V2 challenge presentation. In: 3rd Workshop on Closing the Loop Between Vision and Language (2019)
58. Yu, Y., Ko, H., Choi, J., Kim, G.: End-to-end concept word detection for video captioning, retrieval, and question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
59. Zanfir, M., Marinoiu, E., Sminchisescu, C.: Spatio-temporal attention models for grounded video captioning. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (2016)

60. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
61. Zhou, L., Kalantidis, Y., Chen, X., Corso, J.J., Rohrbach, M.: Grounded video description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6578–6587 (2019)
62. 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 (CVPR), pp. 8739–8748 (2018)