Generating Descriptions with Grounded and Co-Referenced People

Anna Rohrbach[1] Marcus Rohrbach[2] Siyu Tang[1,3] Seong Joon Oh[1] Bernt Schiele[1]
[1]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[2]UC Berkeley, CA, United States [3]Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

Learning how to generate descriptions of images or videos received major interest both in the Computer Vision and Natural Language Processing communities. While a few works have proposed to learn a grounding during the generation process in an unsupervised way (via an attention mechanism), it remains unclear how good the quality of the grounding is and whether it benefits the description quality. In this work we propose a movie description model which learns to generate description and jointly ground (localize) the mentioned characters as well as do visual co-reference resolution between pairs of consecutive sentences/clips. We also propose to use weak localization supervision through character mentions provided in movie descriptions to learn the character grounding. At training time, we first learn how to localize characters by relating their visual appearance to mentions in the descriptions via a semi-supervised approach. We then provide this (noisy) supervision into our description model which greatly improves its performance. Our proposed description model improves over prior work w.r.t. generated description quality and additionally provides grounding and local co-reference resolution. We evaluate it on the MPII Movie Description dataset using automatic and human evaluation measures and using our newly collected grounding and co-reference data for characters.

1. Introduction

When humans talk about what they see, they not only use common objects and terms, but typically refer to reappearing entities, most commonly using names (“John”) and referential words such as pronouns (“he”, “it”). To correctly generate descriptions with reappearing entities, one needs to understand and link them across sentences and visual appearances (images/frames). Current image/video captioning datasets essentially ignore this aspect as they ask to independently describe each image/clip with a single sentence. At the same time, e.g. visual storytelling [17] and movie description [37] ultimately require solving this problem. However, the first approaches on visual storytelling [17] so far have not taken it into account, and current movie description challenges and approaches [35, 48] abstract from it by looking at a single clip at a time and replacing all the character mentions with e.g. “Someone”.

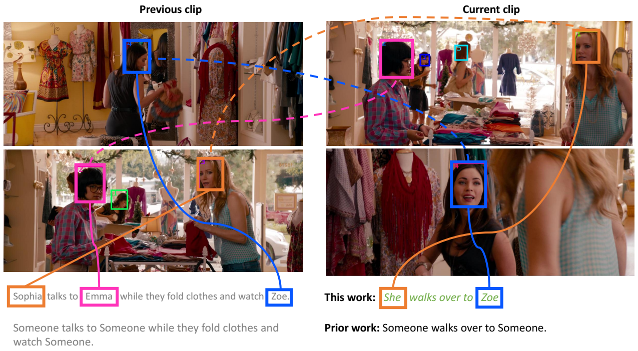
Previous clip
Current clip
Sophia talks to Emma while they fold clothes and watch Zoe.
Someone talks to Someone while they fold clothes and watch Someone.
This work: She walks over to Zoe
Prior work: Someone walks over to Someone.

Figure 1: Bring in the color: our task is to generate grounded and co-referenced descriptions for the current clip using pronouns and new or reappearing character IDs, which are grounded, i.e. localized in the current clip (boxes and lines) and visually co-referenced to the previous clip (dashed lines). The visual grounding allows for co-reference to the previous clip/sentence which enables us using the pronoun “she” to refer to the first ID (Sophia).

In this work we address grounded co-reference resolution, with application to movie description. The most prominent entities in movies are the people or *characters*. In fact, there is a long line of work which aims to link character mentions in movie or TV scripts with their visual tracks [5, 8, 42, 46, 27, 3, 31]. However, all these works are already given the description for all movies where they want to predict the linking. In contrast we want to generate a description, while jointly linking it with the currently and previously depicted character's visual presence. Specifically, the task we address in this work is to *generate descriptions* for movies and at the same time localize or *ground* the characters, recognize their gender and refer to them consistently, i.e. *co-reference* them across sentences, as visualized in Figure 1. Importantly, rather than trying to obtain consistent ids in the entire movie, we focus on robust *local* co-reference resolution on *two consecutive sentences/clips*. We argue that local co-reference resolution is an important problem on itself. On the one hand there are many characters without proper names and/or with only a few occurrences, which can and should be resolved locally, e.g. "*The priest* takes their vows. *He* declares them wife and husband". On the other hand, there are many hard decisions which have to be made locally, e.g. which character to describe and whether a character should be referenced by proper name or pronoun. To clarify, we do not generate the true proper names of the characters, but only identities with gender. We use a predefined set of names in our examples (e.g. Sophia). In future work we believe the true names could be extracted either from dialog, or from one/a few annotations per character.

Approaching the joint description and grounding task requires three main ingredients: we need to *localize the characters*, we need to decide which character(s) to *pay attention to*, and we need to *co-reference* visual characters' appearances in neighboring sentences/clips. In Section 4 we detail how we approach *character localization* using head detection and tracking via a two-stage clustering approach. While generating the sentence, we advocate to *jointly* decide which character to *pay attention to* and if and how to *co-reference* it to the previous grounded characters. In Section 5, we propose to adapt the attention mechanism [1, 53] for this and extend it to attend *jointly* over both problems: grounding (i.e. track selection) and co-reference (i.e. track linking). A key insight is that this can not be learned purely from sentence supervision for generation. Instead, we supervise the joint-attention mechanism with automatically obtained linking of character mentions and tracks (Section 5.2). We note that at test time this supervision is not available and the system has learned, how to jointly ground, co-reference, and describe.

The contributions of our paper include: a) a new task of movie description with grounded and co-referenced characters; to foster research in this direction we will share our newly collected co-reference annotations and grounding of character mentions in the MPII-MD dataset (Section 3); b) a novel approach which addresses this problem by jointly learning to ground the described characters and perform local co-reference resolution between the neighboring clips; c) a robust automatic way of obtaining linking between character mentions in text and visual tracks in video, which we use to supervise our description approach and which we show is essential for the co-reference resolution task.

2. Related Work

Our work aims to do three tasks jointly: generating video descriptions, grounding, and co-reference resolution. We review related work in these three directions with a focus on works which attempt multiple tasks at once. As we focus on people grounding and co-reference, we also discuss the related work on person re-identification and track naming.

Description generation. Generating natural language about visual content has received large interest since the emergence of recurrent networks. Typically the focus is to generate a single sentence about a single image [7, 19, 26, 50, 53], video [7, 12, 33, 38, 47], or most closely to this work, movie clip [34, 49]. Several works also produce grounding while generating the description: [53] propose an attention mechanism to ground each word to spatial CNN image features, [55] extend this to bounding boxes, [54] to video frames, and [59] to spatial-temporal proposals. [24] look into evaluating attention correctness for image captioning. [18] take a different direction and build a model which describes the entire image by jointly predicting large number of bounding boxes and a corresponding short phrase for each box. [23] parse the visual 3D scene and generate coherent multi-sentence descriptions where the objects are grounded in 3D cuboids. Multi-sentence image/video description has also been explored in e.g. [17, 33, 40, 57].

Grounding objects in images/video. Grounding nouns as well as complex natural language expressions in images [16, 20, 25, 30, 32, 52, 58] and video [22, 56] has recently received increased interest. The focus in our work is to localize people in a video while mentioning them in a generated sentence. For example, when mentioning a character who is jogging in a park, we want to localize this person in the video. Additionally we are interested in obtaining visual tracks for character mentions in text, for which rely on the semi-supervised grounding approach from [32].

Co-reference resolution. Co-reference resolution is the task defined in linguistic community [2], where the goal is to establish correct links between named entities and references to them, e.g. pronouns. [31] address co-reference resolution in TV show descriptions with a bidirectional optimization using character visual appearance and linguistic co-reference resolution features.

Person re-identification. Person re-identification from face/head images is a well studied problem and recently many deep learning based approaches have been proposed to address it [21, 28, 39, 43, 44, 62]. Our work is related to this line of work as we aim to re-identify characters between two video clips while generating a video description.

Linking tracks to names. Related works [5, 8, 31, 42, 46] propose datasets for character identification targeting TV shows, which rely on alignment of video to TV scripts. The goal is to track faces in the video and assign names to them. Typically the tracks include background characters. [3] attack the problem of learning a joint model of actors and actions in movies using weak supervision provided by scripts. [27] propose a multiple instance learning based approach which focuses on recognizing background characters, and show significant improvement over prior work. There are two differences between ours and these prior works. First, we aim to re-identify characters locally, without ever seeing them before. Second, when obtaining the matching between names and tracks, our goal is to predict the grounding for a given character, not to name all the tracks.

	Names	Pronouns	All Mentions	Boxes
Training	37,432	15,093	52,525	489
Validation	3,440	1,092	4,532	412
Test	4,453	1,654	6,107	1,748
Total	45,325	17,839	63,164	2,649

Table 1: Left: number of annotated mentions, right: number of named bounding boxes, on MPII-MD [35].

3. A Dataset for Grounded and Co-Referenced Characters

One of the goals in this work is to learn the visual co-reference resolution. To address and evaluate this task we require annotations on language and visual side. On the language side we want to know when different mentions refer to the same person. On the visual side we require grounding of names to visual appearances. Thus we collect new annotations for character co-reference resolution and grounding for the MPII Movie Description (MPII-MD) dataset [35].

Co-reference annotations for character mentions. First, we aim to label all the character mentions in the movie descriptions of the MPII-MD. We semi-automatically[1] annotate names and co-references for each movie. E.g. there might be different ways of referring to the same character ("Mary Jane" as "MJ"), so we link them together under one "alias". Additionally, we annotate the gender of all the characters. As the last step, we annotate pronouns "he" and "she" in all descriptions. When possible we link them to one of the existing names (with some exceptions for rare characters which were not named). In total we label 45,325 name mentions and 17,839 pronouns, see Table 1. With this information we create our corpus **MPII-MD Co-ref+Gender**, where we transform the original MPII-MD descriptions so that every character mention, which appears in a previous sentence, is replaced with "MaleCoref"/"FemaleCoref", otherwise with "MaleName"/"FemaleName". We emphasize that this is the only difference to the standard MPII-MD, i.e. the video clips and splits are identical.

Grounded character annotations. To evaluate the correctness of character grounding we annotate some characters with bounding boxes in video frames. For a subset of movies from MPII-MD Training, Validation and Test set we randomly select sentences and annotate all the mentioned characters. Specifically, whenever the character is mentioned in the sentence and is visible in the corresponding clip, we annotate a few frames of the clip with his/her head bounding boxes. As we also want to evaluate the co-reference correctness, we additionally annotate pairs of consecutive sentences/clips from the Test set. In total we label 2,649 bounding boxes with names, see Table 1.

4. Visual Representations for Characters and their Context

In this section our goal is to localize individual characters in video and extract visual representations informative of their appearance and context. Towards this goal we first detect, track, and extract localized representations for individual characters (Section 4.1), and then extract global representations which capture the scene and context not captured in localized representations (Section 4.2).

4.1. Character tracks and representations

To localize the characters in movies we focus on localizing their heads as most of the time the head of a character is shown, but frequently not the full body. In contrast to prior work [31] we do not only focus on frontal faces but also allow for more challenging, e.g., back views. We detect the heads and track them with a two-step clustering approach, which is able to track across shot boundaries. We extract visual representations on the tracks, informative for estimating characters' identity, activity, gender, and importance.

Head detection. We detect all people in our videos using a head detector. Unlike conventional face detectors, our head detector can reliably detect profile faces and even back view heads. This is desirable as movies contain a large variety of view angles on heads. Our detector is based on the Faster R-CNN [9]. For training we collect head bounding box annotations over the PASCAL VOC 2010 trainval set. The dataset consists of 10,103 images of 7,372 head instances. 6,659 images do not have people, but we retain them as source of negatives. We run our detector[1] on every frame of MPII-MD and keep all the head detections with scores ≥ 0.5 and both dimensions ≥ 40 pixels.

Head tracking. After obtaining the head detections we have to track them within the video clip, i.e. group all detections corresponding to one person together. We have to consider that movies have shot boundaries (rapid changes in a camera viewpoint/angle), thus motion can not be the only cue for tracking, and we require appearance as well. This motivates our two-step approach, where we first group head detections within shots based on their motion and then group the obtained tracks based on their appearance.

We first obtain the shot boundaries with a shot boundary detector[1]. We select the parameters on a set of annotated frames and get the F-score 0.98. We try to detect all boundaries if possible and not produce too many false positives (wrong boundaries). Our tracking framework is based on [45], a multicut [4, 11] tracker for pedestrians in street scene videos. The idea is to build a graph based on person detections in video, and then obtain the tracks by partitioning the graph into an optimal number of connected components, based on attractive and repulsive pairwise terms between pairs of detections. We adapt the multicut tracker

[1] More details can be found in the arXiv version of this paper [36].

to generate tracks for person heads in video clips. We cast our task as a two-level clustering problem. First, we generate tracks from detections that are obtained on the consecutive frames within shots. For that we employ simple geometric features between detection bounding boxes. Given two bounding boxes b and b', with spatial-temporal locations (x, y, t) , scales h and corresponding image regions B, we define the following variables: $\bar{h} = \frac{(h_b+h_{b'})}{2}$, $\Delta x = \frac{|x_b-x_{b'}|}{\bar{h}}$, $\Delta y = \frac{|y_b-y_{b'}|}{\bar{h}}$, $\Delta h = \frac{|h_b-h_{b'}|}{\bar{h}}$, $IOU = \frac{|B_d \cap B_{d'}|}{|B_d \cup B_{d'}|}$, where IOU is the intersection over union of the two detection bounding boxes. The pairwise feature is defined as $(\Delta x, \Delta y, \Delta h, IOU)$. We also add the quadratic terms of each feature to form a nonlinear mapping from feature space to the pairwise potentials. Second, we cluster the obtained tracks, selecting the ones which are at least 5 frames long for computational efficiency. Here we rely on the visual appearance features. For each track we mean pool the FaceVGG [28] fc7 representations on the head crops. We then compute the *cosine* distance between pairs of tracks and use $1-$ distance as pairwise potentials.

Track representations. For re-identification of characters we rely on the FaceVGG [28] fc7 representation, referred to as v^{head} in the following. We mean pool the representations over the head crops clustered together in a track t and refer to it as $v^{head}(t)$. We discuss in Section 5 how we estimate the similarity of two tracks for character re-identification in our pipeline. We include the person body context which could be useful to e.g. predict the person's activity. We extract the body region w.r.t. the head bounding box: 3 times wider and 6 times taller. We experiment with two visual features on the body region. First is the VGG [41] fc7 representation fine-tuned for 393 activities from the *MPII Human Pose Activity* dataset [29], provided by [10]. We only use the body crops ignoring the additional context features as they would be similar across tracks and thus likely not help to distinguish tracks, but significantly increase computation. Another feature is the ResNet [13] (pool5), trained on ImageNet [6] for object classification. We mean pool both visual representations over the body crops in a track and refer to that as $v^{body}(t)$. In the experiments we specify if/which feature is being used. We find, as [27], that the described characters are often in the front, center, and large compared to the background characters. Rather than manually defining a good function we provide the following track statistics $v^{stat}(t)$ and allow our approach to learn from this data: track length, mean and standard deviation of head width/height/center/detection score. We do not extract designated gender features, as we find that v^{head} and v^{body} carry strong information about this aspect. It is straightforward to include a targeted representation as part of future work. All the representations are normalized element-wise by mean centering and dividing by the standard deviation to improve learning subsequent functions with deep learning.

4.2. Holistic video representations

In the previous section we discussed how and which localized features we extract for characters. To additionally capture context, objects, and scene information, important for movie description, we additionally rely on global representations provided by [34] for the MPII-MD dataset. We shortly review them in the following: 1) scores from 146 activity classifiers trained with Dense Trajectory features [51]; 2) scores from 99 object classifiers trained with LSDA [15] responses; 3) scores from 18 scene classifiers trained with PLACES-CNN [61] responses. All the classifiers were trained in [34] using the words from descriptions as labels. The provided visual feature v^{global} is a 263 dimensional concatenation of all three groups of scores.

5. Generating Grounded and Co-Referenced Descriptions

As discussed in the introduction, we focus on character grounding and local co-reference resolution, while generating the description. More specifically, we aim to predict the character grounding and do co-reference resolution given the previous sentence grounding. At test time this allows to e.g. process the movie sequentially from start to end. In the following we rely on our transformed description corpus, **MPII-MD Co-ref+Gender**, described in Section 3.

The key ideas of our approach are to predict grounding and co-reference resolution *jointly* while generating the sentence (Section 5.1) and to learn grounding and co-reference with noisy supervision at training time obtained automatically by linking character mentions and tracks (Section 5.2). Figure 2 provides an overview of our model.

5.1. Predicting grounding and co-reference during sentence generation

For generating sentences we rely on a recurrent LSTM [14] network as defined in [60]. To predict the hidden state at step τ of the sentence, we provide it with the previous word $w_{\tau-1}$ and hidden state $h_{\tau-1}$, as well as the current visual representation v_τ: $h_\tau = f^{LSTM}([w_{\tau-1}, v_\tau], h_{\tau-1})$ where $[,]$ denotes concatenation. The f^{LSTM} has an additional hidden state or memory cell c_t which is not exposed. The word is then predicted as $w_\tau = f^{pred}(h_\tau) = Softmax(W^{pred}h_\tau + b^{pred})$ which can be supervised with the ground truth word $\hat{w}_\tau$. Note that our vocabulary $w \in V$ does not contain any character names, but only $V^{person} = \{\mathit{MaleCoref, FemaleCoref, MaleName, FemaleName}\} \subset V$.

In the following we discuss how we obtain a v_τ which allows to predict the correct word and at the same time solve the grounding and co-reference problem. We formulate the problem in terms of tracks which are the result of the head tracking in Section 4.1. We have tracks $t_c \in T^c$ in the **c**urrent clip $(C = |T^c|)$, and tracks $t_p \in T^p$ in the **p**revious

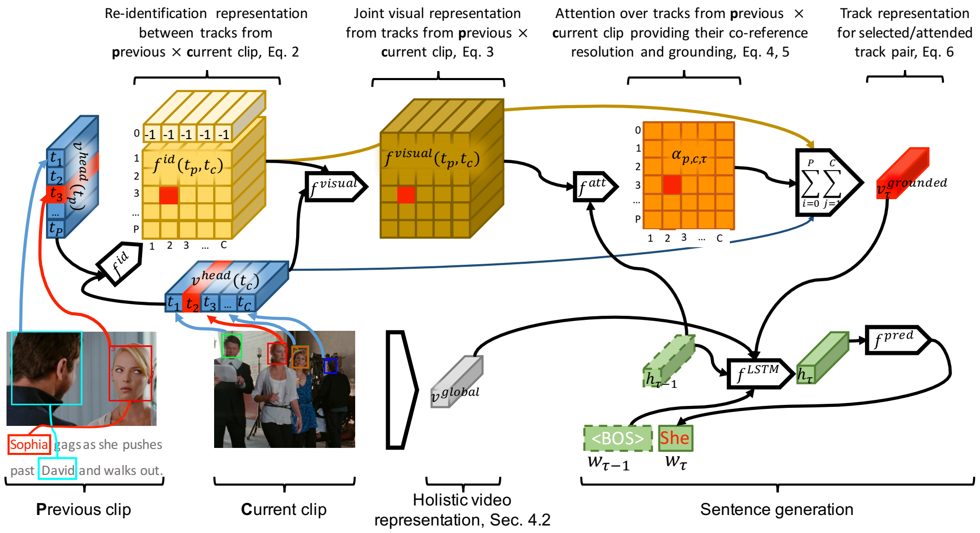
Re-identification representation between tracks from previous × current clip, Eq. 2
Joint visual representation from tracks from previous × current clip, Eq. 3
Attention over tracks from previous × current clip providing their co-reference resolution and grounding, Eq. 4, 5
Track representation for selected/attended track pair, Eq. 6
$v^{head}(t_p)$
t_1 t_2 t_3 ... t_P
$f^{id}(t_p,t_c)$
0 -1 -1 -1 -1 -1
1 2 3 ... P
1 2 3 ... C
f^{visual}
$f^{visual}(t_p,t_c)$
f^{att}
$\alpha_{p,c,\tau}$
$\sum_{i=0}^{P}\sum_{j=1}^{C}$
$v^{grounded}$
f^{id}
$v^{head}(t_c)$
t_1 t_2 t_3 ... t_C
v^{global}
$h_{\tau-1}$
f^{LSTM}
h_τ
f^{pred}
<BOS>
She
$w_{\tau-1}$
w_τ
Sophia gags as she pushes past David and walks out.
Previous clip
Current clip
Holistic video representation, Sec. 4.2
Sentence generation

Figure 2: Our model. Some components are omitted for clarity, e.g. we omit the body and statistic representations.

clip ($P = |T^p|$). We always assume the sentences in the previous clip are already grounded to tracks and only consider those tracks which correspond to mentions of characters in the sentence. Whenever we generate a word w_τ which refers to a person $w_\tau \in V^{person}$, the task is to also select which track $t_{\hat{c}}$ it corresponds to in the current clip and which track $t_{\hat{p}}$ in the previous clip. To account for the case when the person was not mentioned in the previous sentence we include t_0 in T^p which represents a `null` track, which has to be selected to indicate that we describe a new name. As we are only modeling two consecutive clips at a time, this means if $t_{\hat{p}} = t_0$ we want to generate *MaleName* or *FemaleName* and *MaleCoref* or *FemaleCoref* otherwise.

Track re-identification for visual co-reference. To estimate similarity of two tracks t_p and t_c we learn a weighting after element-wise multiplication[2]:

$$v^{id}(t_p,t_c) = v^{head}(t_p) \odot v^{head}(t_c) \tag{1}$$

$$f^{id}(t_p,t_c) = W^{id} v^{id}(t_p,t_c) \tag{2}$$

For $p = 0$, which indicates that no similar track exists, we set $v^{id}(t_0,t_c) = -1$. In preliminary experiments we found that this works better than 0, as values v^{id} are close to 0.

Learning grounding and co-reference jointly. The goal of our approach is to select a track $t_{\hat{c}}$ and the corresponding previous track $t_{\hat{p}}$ which matches the person we are describing with the current word at time τ, in other words we ground this person in $t_{\hat{c}}$ and link it to $t_{\hat{p}}$. As noted above if $t_{\hat{p}} = t_0$ there is no previous track with the same identity as $t_{\hat{c}}$. We propose to jointly predict $\hat{c}$ and $\hat{p}$ using an attention mechanism which takes into account the re-identification and visual representations as well as the hidden state $h_{\tau-1}$ of the recurrent LSTM network generating the description.

The visual features are jointly embedded in the same space as the embedding learned for the hidden state:

$$\begin{aligned} f^{visual}(t_p,t_c) = W^{head} v^{head}(t_c) + W^{body} v^{body}(t_c) \\ + W^{stat} v^{stat}(t_c) + f^{id}(t_p,t_c) + b^v \end{aligned} \tag{3}$$

Afterwards visual and hidden state representation are element-wise multiplied and we learn a function to predict the attention α. This is inspired by [53], who combine convolutional visual features and the recurrent hidden state in the same way to predict spatial attention. Conceptually different, we predict two aspects jointly, the grounding t_p and linking t_c of tracks from different clips.

$$\begin{aligned} \bar{\alpha}_{p,c,\tau} = f^{att}(t_p,t_c,\tau) = \\ W^{\alpha}\phi(W^h h_{\tau-1} + b^h) \odot \phi(f^{visual}(t_p,t_c)) + b^{\alpha} \end{aligned} \tag{4}$$

with the $htan$ non-linearity $\phi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. The attention is normalized with softmax and then we use the predicted α in a weighted sum to get the new local visual representation:

$$\alpha_{p,c,\tau} = \frac{\exp(\bar{\alpha}_{p,c,\tau})}{\sum_{i=0}^{P}\sum_{j=1}^{C}\exp(\bar{\alpha}_{i,k,\tau})} \tag{5}$$

$$\begin{aligned} v_\tau^{grounded} = \sum_{i=0}^{P}\sum_{j=1}^{C} \alpha_{p,c,\tau}[v^{head}(t_c), \\ v^{body}(t_c), v^{stat}(t_c), v^{id}(t_p,t_c)]. \end{aligned} \tag{6}$$

We use this together with the global/holistic video representation v^{global} (see Section 4.2) and the previous word $w_{\tau-1}$ to predict the next hidden state of the recurrent LSTM network as discussed above: $h_\tau = f^{LSTM}([v^{grounded}, v^{global}, w_{\tau-1}], h_{\tau-1})$.

[2] Superscript denotes names of variables/functions, subscript denotes indexes. W / b are learned multiplicative weights / additive bias weights.

Supervising grounding and co-reference. While this system can be trained by only providing reference sentences as supervision, it is difficult to jointly correctly learn the grounding and co-reference resolution. We thus discuss in the next section how to obtain supervision for $\alpha_{p,c,\tau}$. Instead of annotating all characters mentions with tracks, we try to automatically predict the correct track t for each character mention w_τ in the sentence. As we have ground truth co-reference on the text side for the entire training data (Section 3), we can construct the joint ground truth $\hat{\alpha}_{p,c,\tau}$ from the groundings per clip $\hat{\alpha}_{p,\tau}$, $\hat{\alpha}_{c,\tau}$. For all non-character words $w_\tau \notin V^{person}$, no supervision and thus no loss is provided. The losses from sentence supervision and grounding/co-reference supervision are weighted equally.

5.2. Obtaining automatic supervision: linking character mentions and tracks

In this section we discuss how to ground or link character mention with id m_τ in text at position τ to a corresponding visual track t_c in the video to provide ground truth $\hat{\alpha}_{c,\tau}$ used above. In contrast to sentence generation, here we explicitly use the character mentions m (e.g. ”Harry”) which appear in the text. In other words we want to robustly choose the correct track for all character mentions. Note, that this is a slightly different task than in e.g. [27], who aim to link all the visual tracks to correct names. To link the name mentions in text to tracks we adapt the recently proposed semi-supervised approach GroundeR [32]. This approach was initially proposed for the task of localizing text phrases within an image without localization supervision, i.e. where the phrase is located. The main idea is to learn to *attend to* the right bounding box out of a set of proposals, by trying to reconstruct the phrase. We adapt this to our scenario by learning to localize a character $m_{\tau,k}$ in the set of tracks T_k from clip k, where character m is mentioned in the sentence k at position τ. We represent tracks with $v^{head}(t_{c,k})$ and encode character names m together with an identifier of the $gender(m) \in \{M, F\}$ as separate word in an LSTM. Adding the gender allows the model to exploit correlations with different visual appearance of male versus female people and thus helps selecting the right track. In the special case when the sentence k only contains a single name and the clip k contains a single track, i.e. $|T_k| = 1$, we assume that grounding is correct and this information is used as additional supervision, thus enabling the semi-supervised setting of [32]. To train the model we use pairs $([gender(m_{\tau,k}), m_{\tau,k}], \{v^{head}(t_{c,k})\}_{c\in\{1..C\}})$ and predict the grounding as the track with maximum attention from all the tracks in the clip.

6. Evaluation

We first evaluate the quality of our person head detection, tracking, and automatic linking between character names and tracks, obtained in Section 5.2. Then we evaluate the grounded movie description by breaking it down into evaluating description quality and grounding quality.

Recall	Training	Val	Test
Detection	82.00	65.78	84.73
Tracking	78.53	61.65	81.41

Accuracy	Train	Val	Test
GroundeR	78.12	84.46	80.35

Table 2: Left: detection and tracking recall on the annotated heads. Right: linking accuracy on the annotated names/bounding boxes (evaluated on the boxes covered by the tracks). In %.

6.1. Head detection and tracking

We evaluate our head detections and tracks on the collected bounding box annotations from Section 3. Given the annotated bounding boxes we compute detection recall by looking whether there is a head detection in a given frame that has an Intersection Over Union (IOU) ≥ 0.5 with the annotated head box. The track recall is computed similarly, based on the presence of the track that goes through the given frame while overlapping with the annotated box with IOU ≥ 0.5. Table 2(left) shows recall on the Training, Validation and Test parts of the annotations. We analyze the performance of our head detector on the Training annotations and find that multiple factors, e.g. motion blur, occlusion and head size (both small and large), contribute to the missing recall. On the well visible heads we achieve 93.2% recall. The tracking recall is slightly lower, due to the short track rejection (Section 4.1). Tracking can be hard when heads are observed from unusual angles. Overall, we find that our annotations are rather challenging but the obtained performance is reasonable. We also note that our approach already works with just one good track for each character.

6.2. Linking characters to tracks

For every clip we restrict the number of tracks to 50. If $>$ 50 tracks are available we keep the longest, otherwise we zero-complete the missing tracks. For the previous track we consider at most 7 tracks in addition to the “null” track (no match among the previous tracks). Thus there are 8×50 possible choices to predict the character grounding and co-reference during sentence generation. We first train the GroundeR [32] approach on Training movies only to estimate the hyper parameters. Next we combine the Training, Validation and Test movies and train GroundeR on this joint set. We evaluate the accuracy of the obtained predictions on the annotated name/bounding box pairs (Section 3). Given a name we choose the top scoring track as the grounding prediction. For this track we then check whether it contains the annotated frame and overlaps with the annotated box with IOU $\geq$ 0.5. Table 2(right) shows that GroundeR robustly predicts the correct track for a given character name.

6.3. Evaluating description quality

We evaluate our approach in terms of description quality and compare it to a few baselines as well as prior work via automatic and human evaluation. We report all the standard automatic measures in Table 3. The human judges were provided with pairs (reference sentence; predicted sentence), and asked to compare them w.r.t. being helpful for a blind person to follow the events in the video [37]. The judges could decide that one sentence is better than the other or both are similar. Each pair is evaluated by 3 human judges. Next, for every system we compute the percentage of times when at least 2 out of 3 judges decided that the predicted sentence is similar or better than the reference. Table 3 presents the results of human evaluation in the last column.

The top of Table 3 contains the reference numbers from prior works on the standard MPII-MD. We can not use attention supervision or evaluate grounding on standard MPII-MD, which are our core contributions. Our reduced model "Our w/o α" achieves similar scores to prior work.

The middle and bottom part of the table presents results on MPII-MD Co-ref+Gender, thus the numbers between the two settings are not directly comparable, as the references change which strongly affects the automatic measures. To address this we evaluate the approach Visual-Labels [34] on our corpus. Unlike [34], we do not ensemble multiple models. For a fair comparison with the Visual-Labels, in the middle part of Table 3 we provide variants of our model that do not have access to the previous clip character grounding but instead select the 7 biggest previous tracks if sorted by track length times average track area. We compare a variant without the body context features ("Our"), with body features ("Our + Activity") as described in Section 4.1, and one without the attention mechanism but with activity feature encoded jointly with the holistic feature ("Our + Activity w/o attention & co-reference"). In the bottom part of Table 3 we use the automatically obtained previous clip grounding (via Section 5.2, which has access to the previous ground-truth sentence). We compare "Our", "Our+Activity", and "Our+ResNet", and ablate the impact of the grounding and co-reference supervision ("Our w/o $\hat{\alpha}$") and statistic features ("Our w/o statistic features").

From Table 3 we see that: a) the systems "Our" / "Our + Activity" without previous clip character grounding achieve similar or better sentence quality than the Visual-Labels baseline; b) the variant with extra body context but without attention mechanism gets lower human score than our full system (11.0 vs. 15.0); c) providing grounding and co-reference supervision $\hat{\alpha}$ benefits the sentence quality; d) overall, body context features improve the scores, while the statistic features do not have a significant impact; e) the best result, according to human evaluation, is achieved by the variant of our approach "Our + Activity" *without previous clip character grounding*. A possible explanation for this is as follows. In the automatically obtained previous clip's character grounding we might: a) link the characters to tracks correctly; b) link them incorrectly; c) miss some links if names are absent. In a) we follow the storyline of the movie. When we instead use the 7 largest tracks of the previous clip, we bias the description of the current clip differently, e.g. focus on the most salient characters. The obtained descriptions could be ranked higher by the humans, as they only see the current clip in isolation. For b) and c) it is naturally more difficult to get a correct description.

	Automatic				Human
Approach	Bleu-4	Metor	Rouge	CIDEr	judgment
Standard MPII-MD with "Someone"					
Best of [35]	0.47	5.59	13.21	8.14	-
Visual-Labels [34]	0.80	7.03	16.02	9.98	-
S2VT [49]	0.64	7.10	15.69	6.96	-
Our w/o $\hat{\alpha}$	0.84	6.43	16.10	10.66	-
MPII-MD Co-ref+Gender					
without previous clip character grounding					
Visual-Labels (no ensemble)	0.66	5.21	13.94	10.34	11.8
Our + Act. w/o att.&co-ref.	0.74	5.58	14.49	10.22	11.0
Our	0.67	5.06	13.17	10.89	14.8
Our + Activity	0.71	5.31	14.14	11.33	15.0
with previous clip character grounding					
Our w/o $\hat{\alpha}$	0.66	5.82	14.29	10.48	10.8
Our w/o statistic features	0.75	5.81	14.97	11.65	-
Our	0.68	5.81	15.33	11.70	14.0
Our + Activity	0.82	6.17	16.12	12.64	14.5
Our + ResNet	0.88	6.00	15.70	11.76	13.0

Table 3: Automatic and human evaluation of description generation on the test set of MPII-MD; for discussion see Section 6.3.

6.4. Evaluating grounding quality

We evaluate the correctness of the predicted grounding, co-reference and character specific word $w_\tau \in \{$*MaleCoref, FemaleCoref, MaleName, FemaleName*$\}$. We evaluate our predictions on the manually obtained ground-truth (Section 3) and automatically obtained ground-truth (Section 5.2). For each of the named bounding boxes we get the track which overlaps with it most, for every character mention we obtain one or more associated ground-truth tracks. In total we obtain 186 sentences with manually obtained grounding and co-reference. For the automatic annotations we rely on a complete MPII-MD Test set ($6,578$ sentences).

We break down the evaluation in three steps: *Grounding*, *Grounding + Co-Reference*, *Grounding + Co-Reference +* w_τ. We compute precision and recall for each step and report the $F1$ score. Precision is computed as a percentage of predictions $\{\alpha_{p,c,\tau}, w_\tau\}$, which are present in ground-truth. For *Grounding* we check whether the track t_c is present among ground-truth tracks, for *Co-reference* it also has to be correctly linked to a track t_p from the previous clip, for the final step the predicted word w_τ should also be correct. For recall computation we check whether ground-truth pairs $\{\hat{\alpha}_{p,c,\tau}, \hat{w_\tau}\}$ are found among the predictions.

F1 score	manual labeled subset Ground	 +Co-Ref	 $+w_\mathcal{T}$	automatic gt, full set Ground	 +Co-Ref	 $+w_\mathcal{T}$
Baselines with heuristic attention						
[34] Center	59.21	19.33	13.83	36.17	24.52	17.26
[34] LxA	69.58	23.93	18.80	41.62	27.58	19.82
[34] LxA,Sim	69.58	39.05	6.07	41.62	29.76	13.11
Our w/o $\hat{\alpha}$	64.60	21.75	13.47	46.19	28.88	20.41
Our w/o stat.feat.	70.77	50.34	44.57	46.34	38.14	32.87
Our	69.17	53.92	49.55	47.24	38.47	33.88
Our + Activity	71.99	50.54	45.63	53.12	42.15	37.23
Our + ResNet	69.76	51.51	46.54	54.73	43.17	37.92
GroundeR gt	89.10	84.36	84.13			

Table 4: Grounding evaluation on the Test set. For discussion see Section 6.4.

The top part of Table 4 shows a set of baselines where we aim to obtain the grounding and co-reference resolution as a post-processing after the sentence was generated. We use Visual-Labels [34] as a sentence generation baseline. We try multiple heuristics to select the track: central position (Center), size, e.g. length times average area (LxA), and use a simple co-reference resolution method: if there are any tracks in the previous clip, we pick the one which is most similar to the selected track as its co-reference (LxA,Sim). The similarity is estimated as $1 - cosine(v^{head}(t_c), v^{head}(t_p))$. The bottom part of the table lists the variants of our approach introduced earlier. Table 4(left) presents the evaluation with the manually obtained ground-truth. In the bottom line we evaluate the quality of automatic ground-truth predictions from Section 5.2. As we see the predictions are overall quite reliable. Encouraged by that we perform the evaluation on this automatic ground-truth on the complete Test set, Table 4(right). We note, that the manually annotated set covers only 2.8% of the full test set, so the results on the full test set are more stable.

We make the following observations: a) the baselines are competitive in the grounding task, but fall far below our approach in the co-reference task, more pronounced on a full Test set (right); this can be attributed to a more challenging data distribution: the complete Test set contains sentences/clips where people are absent and that has to be recognized correctly, while the manually annotated set always contains people and is biased towards co-references; b) grounding and co-reference supervision $\hat{\alpha}$ is very important to learn the co-reference prediction; c) statistics features, although they do not impact the description quality significantly, benefit the co-reference resolution; d) on the full Test set "Our + Activity" and "Our + ResNet" benefit from additional body context and achieve better performance than the variant "Our"; one observation we make is that these two variants are more accurate with respect to presence/absence of people in the sentence/video which impacts the precision and thus the F1 score; e) our approach is doing quite well in the final task, i.e. the language model correctly learns when to use co-references and recognizes the gender information. In Figure 3 we provide some qualitative examples with the predictions from our approach.

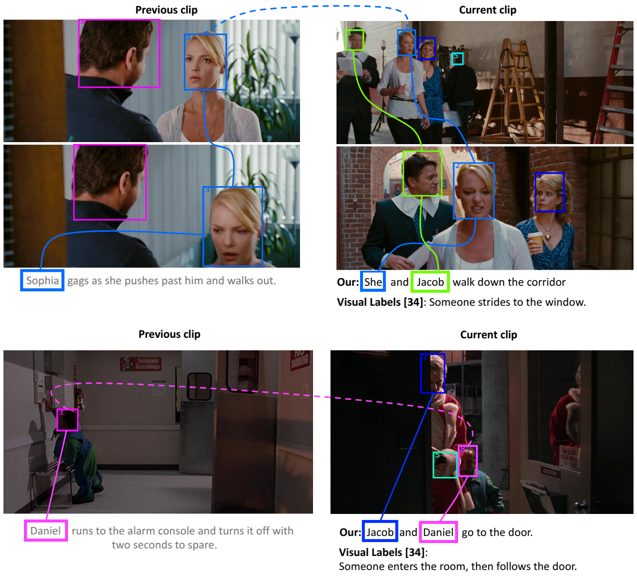
Previous clip
Current clip
Sophia gags as she pushes past him and walks out.
Our: She and Jacob walk down the corridor
Visual Labels [34]: Someone strides to the window.
Previous clip
Current clip
Daniel runs to the alarm console and turns it off with two seconds to spare.
Our: Jacob and Daniel go to the door.
Visual Labels [34]:
Someone enters the room, then follows the door.

Figure 3: Qualitative results of our approach on the grounded movie description task. Given a previous grounding we predict a sentence, grounding and co-reference.

7. Conclusions

In this work we look at the novel task, generating descriptions with joint grounding and co-reference resolution of person mentions. We have proposed a novel approach, which relies on an attention mechanism that jointly learns to solve the grounding and co-reference resolution while learning to describe the video clip. Using an automatically learned linking between names and tracks we can provide supervision into our approach which significantly improves its ability to perform co-reference resolution. We demonstrate encouraging results in a complex task of grounded movie description and achieve improvements over multiple baselines. Our approach generates sentences of better quality than the baselines as shown by automatic and human evaluation. Overall, our approach can describe video, reason about persons identities, recognize their genders and localize them in video. We believe that this work is a first step towards fully coupling generation and grounding while performing image/video description. We will release the annotations and extracted tracks and hope that this will benefit other researchers who work on linguistic and/or visual co-reference resolution, movie question answering, visual storytelling, and multi-sentence video description.

Acknowledgments. We would like to thank Trevor Darrell for helpful discussions. Marcus Rohrbach was supported by the Berkeley Artificial Intelligence Research (BAIR) Lab.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 2

[2] S. Bergsma and D. Lin. Bootstrapping path-based pronoun resolution. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 33–40. Association for Computational Linguistics, 2006. 2

[3] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 1, 2

[4] S. Chopra and M. Rao. The partition problem. *Mathematical Programming*, 59(1–3):87–115, 1993. 3

[5] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1, 2

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 4

[7] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[8] M. Everingham, J. Sivic, and A. Zisserman. ”hello! my name is... buffy” - automatic naming of characters in tv video. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2006. 1, 2

[9] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 3

[10] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1080–1088, 2015. 4

[11] M. Grötschel and Y. Wakabayashi. A cutting plane algorithm for a clustering problem. *Mathematical Programming*, 45(1):59–96, 1989. 3

[12] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shoot recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 2

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[15] J. Hoffman, S. Guadarrama, E. Tzeng, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. LSDA: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 4

[16] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[17] T.-H. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. Visual storytelling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2016. 1, 2

[18] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[19] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[20] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[21] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[22] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2657–2664. IEEE, 2014. 2

[23] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Generating multi-sentence natural language descriptions of indoor scenes. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015. 2

[24] C. Liu, J. Mao, F. Sha, and A. Yuille. Attention correctness in neural image captioning. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2017. 2

[25] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[26] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 2

[27] O. M. Parkhi, E. Rahtu, and A. Zisserman. 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. 1, 2, 4, 6

[28] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015. 2, 4

[29] L. Pishchulin, M. Andriluka, and B. Schiele. Fine-grained activity recognition with holistic and pose based features. In

Proceedings of the German Confeence on Pattern Recognition (GCPR), pages 678–689, 2014. 4

[30] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[31] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking people in videos with "their" names using coreference resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 1, 2, 3

[32] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2, 6

[33] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. In *Proceedings of the German Conference on Pattern Recognition (GCPR)*, 2014. 2

[34] A. Rohrbach, M. Rohrbach, and B. Schiele. The long-short story of movie description. In *Proceedings of the German Conference on Pattern Recognition (GCPR)*, 2015. 2, 4, 7, 8

[35] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 3, 7

[36] A. Rohrbach, M. Rohrbach, S. Tang, S. J. Oh, and B. Schiele. Generating descriptions with grounded and co-referenced people. *arXiv:1704.01518*, 2017. 3

[37] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele. Movie description. *International Journal of Computer Vision (IJCV)*, 2017. 1, 7

[38] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 2

[39] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[40] A. Shin, K. Ohnishi, and T. Harada. Beyond caption to narrative: Video captioning with multiple sentences. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2016. 2

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 4

[42] J. Sivic, M. Everingham, and A. Zisserman. "who are you?"-learning person specific classifiers from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1, 2

[43] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[44] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. 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. 2

[45] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Subgraph decomposition for multi-target tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5033–5041, 2015. 3

[46] M. Tapaswi, M. Baeuml, and R. Stiefelhagen. "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. 1, 2

[47] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2014. 2

[48] A. Torabi, C. Pal, H. Larochelle, and A. Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv:1503.01070v1*, 2015. 1

[49] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 7

[50] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[51] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 4

[52] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[53] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015. 2, 5

[54] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[55] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[56] H. Yu and J. M. Siskind. Grounded language learning from videos described with sentences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013. 2

[57] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[58] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *Proceedings of the*

European Conference on Computer Vision (ECCV), pages 69–85. Springer, 2016. 2

[59] M. Zanfir, E. Marinoiu, and C. Sminchisescu. Spatio-temporal attention models for grounded video captioning. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2016. 2

[60] W. Zaremba and I. Sutskever. Learning to execute. *arXiv preprint arXiv:1410.4615*, 2014. 4

[61] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *Advances in Neural Information Processing Systems (NIPS)*, 2014. 4

[62] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of lfw benchmark or not? *arXiv:1501.04690*, 2015. 2