# Character Grounding and Re-Identification in Story of Videos and Text Descriptions

Youngjae Yu[1,2], Jongseok Kim[1], Heeseung Yun[1],
Jiwan Chung[1], and Gunhee Kim[1,2]

[1] Seoul National University, Seoul, Korea
[2] Ripple AI, Seoul, Korea
{yj.yu,js.kim,heeseung.yun,jiwanchung}@vision.snu.ac.kr,
{gunhee}@snu.ac.kr
http://vision.snu.ac.kr/projects/CharacterReid/

**Abstract.** We address character grounding and re-identification in multiple story-based videos like movies and associated text descriptions. In order to solve these related tasks in a mutually rewarding way, we propose a model named *Character in Story Identification Network* (CiSIN). Our method builds two semantically informative representations via joint training of multiple objectives for character grounding, video/text re-identification and gender prediction: Visual Track Embedding from videos and Textual Character Embedding from text context. These two representations are learned to retain rich semantic multimodal information that enables even simple MLPs to achieve the state-of-the-art performance on the target tasks. More specifically, our CiSIN model achieves the best performance in the Fill-in the Characters task of LSMDC 2019 challenges [35]. Moreover, it outperforms previous state-of-the-art models in M-VAD Names dataset [30] as a benchmark of multimodal character grounding and re-identification.

## 1 Introduction

Searching persons in videos accompanying with free-form natural language descriptions is a challenging problem in computer vision and natural language research [30,32,34,35]. For example, in the story-driven videos such as movies and TV series, distinguishing *who is who* is a prerequisite to understanding the relationships between characters in the storyline. Thanks to the recent rapid progress of deep neural network models for joint visual-language representation [46,42,28,19], it has begun to be an achievable goal to understand interactions of characters that reside in the complicated storyline of videos and associate text.

In this work, we tackle the problem of character grounding and re-identification in consecutive pairs of movie video clips and corresponding language descriptions. The character grounding indicates the task of locating the character mentioned in the text within videos. The re-identification can be done in both text and image domain; it groups the tracks of the same person across video clips or identifies tokens of the identical person across story sentences. As the main

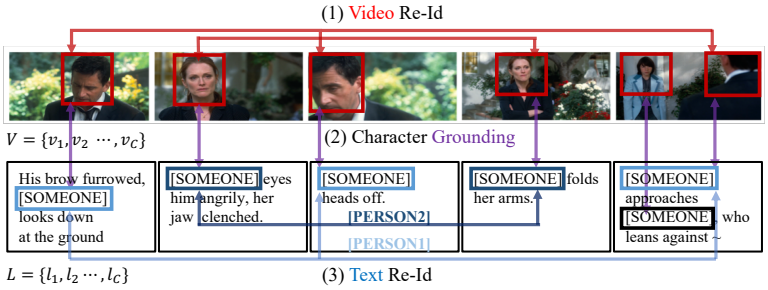


**Fig. 1.** The problem statement. Given consecutive $C$ pairs of video clips and corresponding language descriptions, our CiSIN model aims at solving three multimodal tasks in a mutually rewarding way. Character grounding is the identity matching between person tracks and `[SOMEONE]` tokens. Video/text re-identification is the identity matching between person tracks in videos and `[SOMEONE]` tokens in text, respectively.

testbed of our research, we choose the recently proposed Fill-in the Characters task of the Large Scale Movie Description Challenge (LSMDC) 2019 [35], since it is one of the most large-scale and challenging datasets for character matching in videos and associated text. Its problem statement is as follows. Given five pairs of video clips and text descriptions that include the `[SOMEONE]` tokens for characters, the goal is to identify which `[SOMEONE]` tokens are identical to one another. This task can be tackled minimally with text re-identification but can be synergic to jointly solve with video re-identification and character grounding.

We propose a new model named *Character-in-Story Identification Network* (CiSIN) to jointly solve the character grounding and video/text re-identification in a mutually rewarding way, as shown in Figure 1. The character grounding, which connects the characters between different modalities, complements both visual and linguistic domains to improve re-identification performance. In addition, each character's grounding can be better solved by closing the loop between both video/text re-identification and neighboring character groundings.

Our method proposes two semantically informative representations. First, Visual Track Embedding (VTE) involves motion, face and body-part information of the tracks from videos. Second, Textual Character Embedding (TCE) learns rich information of characters and their actions from text using BERT [6]. They are trained together to share various multimodal information via multiple objectives, including character grounding, video/text re-identification and attribute prediction. The two representations are powerful enough for simple MLPs to achieve state-of-the-art performance on the target tasks.

We summarize the contributions of this work as follows.

1. We propose the CiSIN model that can jointly tackle character grounding and re-identification in both video and text narratives. To the best of our knowledge, our work is the first to jointly solve these three tasks, each of which has been addressed separately in previous research. Our model is jointly trained

via multi-task objectives of these complementary tasks in order to create synergetic effects from one domain to another and vice versa.
2. Our CiSIN model achieves the best performance so far in two benchmarks datasets: LSMDC 2019 challenge [35] and M-VAD Names [30] dataset. The CiSIN model attains the best accuracy in the Fill-in the Characters task of LSMDC 2019 challenges. Moreover, it achieves the new state-of-the-art results on grounding and re-identification tasks in M-VAD Names.

## 2 Related Work

**Linking characters with visual tracks**. There has been a long line of work that aims to link character names in movie or TV scripts with their corresponding visual tracks [7,38,2,32,29,40,27,14]. However, this line of research deals with more constrained problems than ours; for example, having the templates of main characters or knowing which characters appear in a clip. On the other hand, our task requires person grounding and re-identification in more free-formed multiple sentences, without ever seeing characters before.

**Human retrieval with natural language**. Visual content retrieval with natural language queries has been mainly addressed by joint visual-language embedding models [42,18,12,24,43,13,26,22,21], and extended to the video domain [46,42,28]. Such methodologies have also been applied to movie description datasets such as MPII-MD [34] and M-VAD Names [30]. Pini *et al*. [30] propose a neural network that learns a joint multimodal embedding for human tracking in videos and verb embedding in text. Rohrbach *et al*. [34] develop an attention model that aligns human face regions and mentioned characters in the description. However, most previous approaches tend to focus on the retrieval of human bounding boxes with sentence queries in a single clip while ignoring story coherence. On the other hand, our model understands the context within consecutive videos and sentences; thereby can achieve the best result reported so far in the human retrieval in the video story.

**Person re-identification in multiple videos**. The goal of person re-identification (re-id) is to associate with identical individuals across multiple camera views [9,10,8,23,51,1,4,52,20,47,15]. Some methods exploit part information for better re-identification against occlusion and pose variation [37,39,44]. However, this line of work often has dealt with visual information only and exploited the videos with less diverse human appearances (*e.g.* standing persons taken from CCTV) than movies taken by various camerawork (*e.g.* head shots, half body shots and two shots). Moreover, our work takes a step further by considering the video description context of story-coherent multiple clips for character search and matching.

**Visual co-reference resolution.** Visual co-reference resolves pronouns and linked character mentions in text with the character appearances in video clips. Ramanathan *et al*. [32] address the co-reference resolution in TV show description with the aid of character visual models and linguistic co-reference resolution features. Rohrbach *et al*. [34] aim generate video descriptions with grounded

and co-referenced characters. In the visual dialogue domain [36,19], the visual co-reference problem is addressed to identify the same entity/object instances in an image for answering questions with pronouns. On the other hand, we focus on identifying all character mentions blanked as `[SOMEONE]` tokens in multiple descriptions. This is inherently different from the co-reference resolution that can use pronouns or explicit gender information (*e.g.* she, him) as clues.

## 3 Approach

**Problem Statement**. We address the problem of character identity matching in the movie clips and associated text descriptions. To be specific, we follow the Fill-in the Characters task of the Large Scale Movie Description Challenge (LSMDC) 2019[3]. As shown in Figure 1, the input is a sequence of $C$ pairs of video clips and associated descriptions that may include the `[SOMEONE]` tokens for characters (*i.e.* $C = 5$ in LSMDC 2019). The `[SOMEONE]` tokens correspond to proper nouns or pronouns of characters in the original description. The goal of the task is to replace the `[SOMEONE]` tokens with *local* character IDs that are consistent within the input set of clips, not globally in the whole movie.

We decompose the Fill-in the Characters task into three subtasks: (*i*) *character grounding* finds the character tracks of each `[SOMEONE]` token in the video, (*ii*) *video re-identification* that groups visual tracks of the identical character in videos and (*iii*) *text re-identification* that connects between the `[SOMEONE]` tokens of the same person. We define these three subtasks because each of them is an important research problem with its own applications, and jointly solving the problems is mutually rewarding. Thereby we can achieve the best result reported so far on LSMDC 2019 and M-VAD Names [30] (See the details in section 4).

In the following, we first present how to define person tracks from videos (section 3.1). We then discuss two key representations of our model: Visual Track Embedding (VTE) for person tracks (section 3.2) and Textual Character Embedding (TCE) for `[SOMEONE]` tokens (section 3.3). We then present the details of our approach to character grounding (section 3.4) and video/text re-identification (section 3.5). Finally, we discuss how to train all the components of our model via multiple objectives so that the solution to each subtask helps one another (section 3.6). Figure 2 illustrates the overall architecture of our model.

### 3.1 Video Preprocessing

For each of $C$ video clips, we first resize it to 224×224 and uniformly sample 24 frames per second. We then detect multiple person tracks as basic character instances in videos for grounding and re-identification tasks. That is, the tasks reduce to the matching problems between person tracks and `[SOMEONE]` tokens.

**Person tracks.** We obtain person tracks $\{\mathbf{t}_m\}_{m=1}^{M}$ as follows. $M$ denotes the number of person tracks throughout $C$ video clips. First, we detect bounding

[3] `https://sites.google.com/site/describingmovies/lsmdc-2019`.

boxes (bbox) of human bodies using the rotation robust CenterNet [53] and group them across consecutive frames using the DeepSORT tracker [45]. The largest bbox in each track is regarded as its representative image $\{\mathbf{h}_m\}_{m=1}^{M}$, which can be used as a simple representation of the person track.

**Face regions.** Since person tracks are not enough to distinguish "who is who", we detect the faces for better identification. In every frame, we obtain face regions using MTCNN [49] and resize them to 112×112. We then associate each face detection with the person track that has the highest IoU score.

**Track metadata.** We extract track metadata $\{\mathbf{m}_m\}_{m=1}^{M}$, which include the $(x, y)$ coordinates, the track length and the size of the representative image $\mathbf{h}_m$. All of them are normalized with respect to the original clip size and duration.

### 3.2 Visual Track Embedding

For better video re-identification, it is critical to correctly measure the similarity between track $i$ and track $j$. As rich semantic representation of tracks, we build Visual Track Embedding (VTE) as a set of motion, face and body-part features.

**Motion embedding.** We apply the I3D network [3] to each video clip and obtain the last CONV feature of the final Inception module, whose dimension is (time, width, height, feature) = $(\lfloor \frac{t}{8} \rfloor, 7, 7, 1024)$. That is, each set of 8 frames is represented by a $7 \times 7$ feature map whose dimension is 1024. Then, we extract motion embedding $\{\mathbf{i}_m\}_{m=1}^{M}$ of each track by mean-pooling over the cropped spatio-temporal tensor of this I3D feature.

**Face embedding.** We obtain the face embedding $\{\mathbf{f}_m\}_{m=1}^{M}$ of a track, by applying ArcFace [5] to all face regions of the track and mean-pooling over them.

**Body-part embedding.** To more robustly represent visual tracks against the ill-posed views or cameraworks (*e.g.* no face is shown), we also utilize the body parts of a character in a track (*e.g.* limb and torso). We first extract the pose keypoints using the pose detector of [53], and then obtain the body-part embedding $\{\{\mathbf{p}_{m,k}\}_{k=1}^{K}\}_{m=1}^{M}$ by selecting keypoint-corresponding coordinates from the last CONV layer ($7 \times 7 \times 2048$) of ImageNet pretrained ResNet-50 [11]:

$$\mathbf{p}_{m,k} = \text{ResNet}(\mathbf{h}_m)[x_k, y_k], \tag{1}$$

where $\mathbf{p}_{m,k}$ is the representation of the $k$-th keypoint in a pose, $\mathbf{h}_m$ is the representative image of the track, and $(x_i, y_i)$ is a relative position of the keypoint. To ensure the quality of pose estimation, we only consider keypoints whose confidences are above a certain threshold ($\tau = 0.3$).

In summary, VTE of each track $m$ includes three sets of embeddings for motion, face and body parts: VTE = $\{\mathbf{i}_m, \mathbf{f}_m, \mathbf{p}_m\}$.

### 3.3 Textual Character Embedding

Given $C$ sentences, we make a unified language representation of the characters (*i.e.* `[SOMEONE]` tokens), named Textual Character Embedding (TCE) as follows.

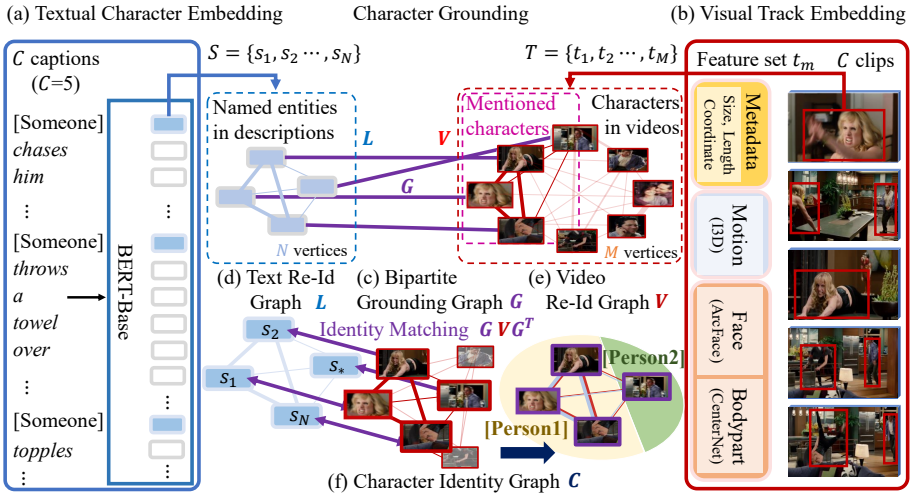


**Fig. 2.** Overview of the proposed *Character-in-Story Identification Network* (CiSIN) model. Using (a) Textual Character Embedding (TCE) (section 3.3) and (b) Visual Track Embedding (VTE) (section 3.2), we obtain the (c) bipartite Character Grounding Graph **G** (section 3.4). We build (d)–(e) Video/Text Re-Id Graph **V** and **L**, from which (f) Character Identity Graph **C** is created. Based on the graphs, we can perform the three subtasks jointly (section 3.5).

**Someone embedding**. We use the BERT model [6] to embed the `[SOMEONE]` tokens in the context of $C$ consecutive sentences. We load pretrained BERT-Base-Uncased with initial weights. We denote each sentence $\{\mathbf{w}_l^c\}_{l=1}^{W_c}$ where $W_c$ is the number of words in the $c$-th sentence. As an input to the BERT model, we concatenate the $C$ sentences $\{\{\mathbf{w}_l^c\}_{l=1}^{W_c}\}_{c=1}^{C}$ by placing the `[SEP]` token in each sentence boundary. We obtain the embedding of each `[SOMEONE]` token as

$$\mathbf{b}_n = \mathrm{BERT}(\{\{\mathbf{w}_l^c\}_{l=1}^{W_c}\}_{c=1}^{C}, k), \tag{2}$$

where $k$ is the position of the `[SOMEONE]` token. We let this word representation $\{\mathbf{b}_n\}_{n=1}^{N}$ (*i.e.* the BERT output at position $k$), where $N$ is the number of `[SOMEONE]` tokens in the $C$ sentences.

**Action embedding**. For a better representation of characters, we also consider the actions of `[SOMEONE]` described in the sentences. We build a dependency tree for each sentence with the Stanford Parser [31], and find the `ROOT` word associated with `[SOMEONE]`, which is generally the verb of the sentence. We then obtain the `ROOT` word representation $\{\mathbf{a}_n\}_{n=1}^{N}$ as done in Eq.(2).

In summary, TCE of each character $n$ includes two sets of embeddings for someone and action. Since they have the same dimension as the output of BERT, we simply concatenate them as a single embedding: TCE $= \mathbf{s}_n = [\mathbf{b}_n; \mathbf{a}_n]$.

### 3.4 Character Grounding

For each clip $c$, we perform the character grounding using VTE $= \{\mathbf{i}_m, \mathbf{f}_m, \mathbf{p}_m\}$ of person tracks $\mathbf{t}_m$ (section 3.2) and TCE $= \mathbf{s}_n = [\mathbf{b}_n; \mathbf{a}_n]$ of characters (section 3.3). Due to the heterogeneous nature of our VTE (*i.e.* $\mathbf{f}_m$ contains appearance information such as facial expression while $\mathbf{i}_m$ does motion information), we separately fuse VTE with TCE and then concatenate to form fused ground representation $\mathbf{g}_{n,m}$:

$$\mathbf{g}^{\text{face}}_{n,m} = \text{MLP}(\mathbf{s}_n) \odot \text{MLP}(\mathbf{f}_m), \quad \mathbf{g}^{\text{motion}}_{n,m} = \text{MLP}(\mathbf{s}_n) \odot \text{MLP}(\mathbf{i}_m), \tag{3}$$

$$\mathbf{g}_{n,m} = [\text{MLP}(\mathbf{m}_m); \mathbf{g}^{\text{face}}_{n,m}; \mathbf{g}^{\text{motion}}_{n,m}], \tag{4}$$

where $\odot$ is a Hadamard product, [;] is concatenation, $\mathbf{m}_m$ is the track metadata, and the MLP is made up of two FC layers with the identical output dimension.

Finally, we calculate the bipartite Grounding Graph $\mathbf{G} \in \mathbb{R}^{N \times M}$ as

$$\mathbf{G}_{nm} = g(\mathbf{s}_n, \mathbf{t}_m) = \text{MLP}(\mathbf{g}_{n,m}), \tag{5}$$

where the MLP consists of two FC layers with the scalar output.

**Attributes.** In addition to visual and textual embeddings, there are some attributes of characters that may be helpful for grounding. For example, as in MPII-MD Co-ref [34], gender information can be an essential factor for matching characters. However, most `[SOMEONE]` tokens have vague context to hardly infer gender information, although some of them may be clear like gendered pronouns (*e.g.* she, he) or nouns (*e.g.* son, girl). Thus, we add an auxiliary attribute module that predicts the gender of `[SOMEONE]` using $\mathbf{b}_n$ as input:

$$attr_{logit} = \sigma(\text{MLP}(\mathbf{b}_n)), \tag{6}$$

where the MLP has two FC layers with scalar output, and $\sigma$ is a sigmoid function.

The predicted attribute is neither explicitly used for any grounding or re-identification tasks. Instead, the module is jointly trained with other components to encourage the shared character representation to learn the gender context implicitly. Moreover, it is straightforward to extend this attribute module to other information beyond gender like the characters' roles.

### 3.5 Re-Identification

We present our method to solve video/text re-identification tasks, and explain how to aggregate all subtasks to achieve the Fill-in the Characters task.

**Video Re-Id Graph.** We calculate the Video Re-Id Graph $\mathbf{V} \in \mathbb{R}^{M \times M}$ that indicates the pairwise similarity scores between person tracks. We use the face and body-part embeddings of VTE. We first calculate the face matching score of track $i$ and $j$, namely $f_{i,j}$, by applying a 2-layer MLP to face embedding $\mathbf{f}_m$ followed by a dot product and a scaling function:

$$f_{i,j} = w_f^{scale} \mathbf{f}'_i \cdot \mathbf{f}'_j + b_f^{scale}, \quad \mathbf{f}'_m = \text{MLP}(\mathbf{f}_m), \tag{7}$$

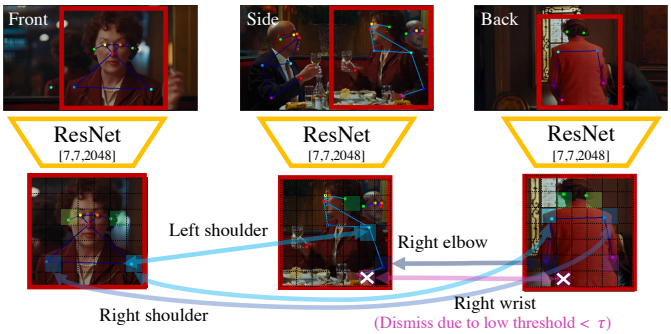


**Fig. 3.** Pose-based body-part matching. We only consider keypoints that appear across scenes with confidences scores above a certain threshold $\tau(= 0.3)$ such as right/left shoulder and right elbow, while ignoring right wrist as it falls short of the threshold.

where $w_f^{scale}, b_f^{scale} \in \mathbb{R}$ are learnable scalar weights for scaling.

We next calculate the pose-guided body-part matching score $q_{i,j}$. Unlike conventional re-identification tasks, movies may not contain full keypoints of a character due to camerawork like upper body close up shots. Therefore, we only consider the pairs of keypoints that are visible in both tracks (Figure 3):

$$q_{i,j} = \frac{w_p^{scale}}{Z_{i,j}} \sum_k \delta_{i,k}\delta_{j,k}\mathbf{p}_{i,k} \cdot \mathbf{p}_{j,k} + b_p^{scale}, \tag{8}$$

where $\mathbf{p}_{i,k}$ is the body-part embedding of VTE for keypoint $k$ of track $i$, $\delta_{i,k}$ is its binary visibility value (1 if visible otherwise 0), $Z_{i,j} = \sum_k \delta_{i,k}\delta_{j,k}$ is a normalizing constant and $w_p^{scale}, b_p^{scale}$ are scalar weights.

Finally, we obtain Video Re-Id Graph $\mathbf{V}$ by summing both face matching score and body-part matching score:

$$\mathbf{V}_{i,j} = f_{i,j} + q_{i,j}. \tag{9}$$

**Text Re-Id Graph.** The Text Re-Id Graph $\mathbf{L} \in \mathbb{R}^{N \times N}$ measure the similarity between every pair of `[SOMEONE]` token as

$$\mathbf{L}_{i,j} = \sigma(\mathrm{MLP}(\mathbf{b}_i \odot \mathbf{b}_j)), \tag{10}$$

where $\odot$ is a Hadamard product, $\sigma$ is a sigmoid and the MLP has two FC layers.

**Solutions to three subtasks.** Since we have computed all pairwise similarity between and within person tracks and `[SOMEONE]` tokens in Eq.(5 and Eq.(9)–(10), we can achieve the three tasks by thresholding. We perform the character grounding by finding the column with the maximum value for each row in the bipartite Grounding Graph $\mathbf{G} \in \mathbb{R}^{N \times M}$. The video re-identification is carried out by finding the pair whose score in Video Re-Id Graph $\mathbf{V}$ is positive. Finally, the text re-identification can be done as will be explained below for the Fill-in the Characters task since they share the same problem setting.

**Fill-in the characters task**. We acquire the Character Identity Graph $\mathbf{C}$:

$$\mathbf{C} = \text{avg}(\mathbf{L}, \mathbf{R}) \quad \text{where } \mathbf{R} = \sigma(\mathbf{G}\mathbf{V}\mathbf{G}^T),\ \mathbf{G}_n = \text{argmax}_{m \in M_c}\, g(\mathbf{s}_n, \mathbf{t}_m), \quad (11)$$

where $\sigma$ is a sigmoid, $\mathbf{G} \in \mathbb{R}^{N \times M}$ is the bipartite Grounding Graph in Eq.(5), $\mathbf{V} \in \mathbb{R}^{M \times M}$ is the Video Re-Id Graph in Eq.(9), and $\mathbf{L} \in \mathbb{R}^{N \times N}$ is the Text Re-Id Graph in Eq.(10). We perform the Fill-in the Characters task by finding $\mathbf{C}_{ij} \geq 0.5$, for which we decide `[SOMEONE]` token $i$ and $j$ as the same character.

Although we can solve the task using only the Text Re-Id graph $\mathbf{L}$, the key idea of Eq.(11) is that we also consider the other loop leveraging character grounding $\mathbf{G}$ and the Video Re-Id graph $\mathbf{V}$. That is, we find the best matching track for each token in the same clip $c$, where $M_c$ is the candidate tracks in the clip $c$. We then use the Video Re-Id graph to match tracks and apply $\mathbf{G}$ again to find their character tokens. Finally, we average the scores from these two loops of similarity between `[SOMEONE]` tokens.

### 3.6 Joint Training

We perform joint training of all components in the CiSIN model so that both VTE and TCE representation share rich multimodal semantic information, and subsequently help solve character grounding and video/text re-identification in a mutually rewarding way. We first introduce the loss functions.

**Losses**. For character grounding, we use a triplet loss where a positive pair maximizes the ground matching score $g$ in Eq.(5) while a negative one minimizes

$$\mathcal{L}(\mathbf{s}, \mathbf{t}, \mathbf{t}^-) = \max(0,\ \alpha - g(\mathbf{s}, \mathbf{t}) + \ g(\mathbf{s}, \mathbf{t}^-)), \quad (12)$$

$$\mathcal{L}(\mathbf{s}, \mathbf{t}, \mathbf{s}^-) = \max(0,\ \alpha - g(\mathbf{s}, \mathbf{t}) + \ g(\mathbf{s}^-, \mathbf{t})), \quad (13)$$

where $\alpha$ is a margin, $(\mathbf{s}, \mathbf{t})$ is a positive pair, $\mathbf{t}^-$ is a negative track, and $\mathbf{s}^-$ is a negative token. For video re-identification, we also use a triplet loss:

$$\mathcal{L}(\mathbf{t}_0, \mathbf{t}_+, \mathbf{t}_-) = \max(0,\ \beta - \mathbf{V}_{0,+} + \mathbf{V}_{0,-}) \quad (14)$$

where $\beta$ is a margin and $\mathbf{V}$ is the score in Video Re-Id Graph. $(\mathbf{t}_0, \mathbf{t}_+)$ and $(\mathbf{t}_0, \mathbf{t}_-)$ are positive and negative track pair, respectively. For text re-identification, we train parameters to make Character Identity Graph $\mathbf{C}$ in Eq.(11) closer to the ground truth $\mathbf{C}^{gt}$ with a binary cross-entropy (BCE) loss. When computing $\mathbf{C}$ in Eq.(11), we replace the argmax of $\mathbf{G}_n$ with the softmax for differentiability: $\mathbf{G}_n = \text{softmax}_{m \in M} g(\mathbf{s}_n, \mathbf{t}_m)$. Additionally, the attribute module is trained with the binary cross-entropy loss for gender class (*i.e.* female, male).

**Training**. We use all losses to train the model jointly. While fixing the parameters in the motion embedding (I3D) and the face embedding (ArcFace), we update all the other parameters in all MLPs, ResNets and BERTs during training. Notably, BERT models in TCE are trained by multiple losses (*i.e.* character grounding, text re-identification and the attribute loss) to learn better multimodal representation.

In Eq.(11), the Grounding Graph **G** is a bipartite graph for cross-domain retrieval between the Text Re-Id Graph **L** and the Video Re-Id Graph **V**. The bipartite graph **G** identifies a subgraph of **V** for the characters mentioned in the text such that the subgraph is topologically similar to **L**. By joint training of multiple losses, the similarity metric between the visual and textual representation of the same character increases, consequently improving both character grounding and re-identification performance.

**Details**. We unify the hidden dimension size of all MLPs as 1024 and use the leaky ReLU activation for every FC layer in our model. The whole model is trained with the Adam optimizer [17] with a learning rate of $10^{-5}$ and a weight decay of $10^{-8}$. We train for 86K iterations for 15 epochs. We set the margin $\alpha = 3.0$ for the triplet loss in Eq.(12)–(13) and $\beta = 2.0$ for Eq.(14).

## 4 Experiments

We evaluate the proposed CiSIN model in two benchmarks of character grounding and re-identification: M-VAD Names [30] dataset and LSMDC 2019 challenge [35], on which we achieve new state-of-the-art performance.

### 4.1 Experimental Setup

**M-VAD Names.** We experiment three tasks with M-VAD Names [30] dataset: character grounding and video/text re-identification. We group five successive clips into a single set, which is the same setting with LSMDC 2019, to evaluate the model's capability to understand the story in multiple videos and text. M-VAD Names dataset is annotated with persons' name and their face tracks, which we use as ground truth for training.

For evaluation of character grounding, we measure the accuracy by checking whether the positive track is correctly identified. For evaluation of video re-identification, we predict the Video Re-Id Graph and then calculate the accuracy by comparing it with the ground truth. For evaluation of text re-identification, we calculate the accuracy with the ground truth character matching graph.

**LSMDC 2019.** We report experimental results for the Fill-in the Characters task of LSMDC 2019 [35], which is the superset of M-VAD dataset [41]. Contrary to M-VAD Names, LSMDC 2019 only provides local ID ground truths of `[SOMEONE]` tokens with no other annotation (*e.g.* face bbox annotation and characters' names). We exactly follow the evaluation protocol of the challenge.

### 4.2 Quantitative Results

**M-VAD Names.** Table 1 summarizes the quantitative results of video re-identification tasks. We compare with state-of-the-art strong-ReID-baseline [25] using the official implementation[4] which is pretrained on market-1501 [51]. For

[4] `https://github.com/michuanhaohao/reid-strong-baseline`.

**Table 1.** Results of video re-identification on the validation set of M-VAD Names [30].

| Video Re-Id | Accuracy |
|---|---|
| ResNet-50 (ImageNet pretrained) | 0.607 |
| strong-ReID-baseline [25] | 0.617 |
| Face+fullbody+poselets [50] | 0.776 |
| Face+head+body+upperbody [16] | 0.779 |
| RANet visual context [15] | 0.787 |
| CiSIN Face only | 0.783 |
| + Human bbox | 0.781 |
| + Body parts | 0.799 |
| + Body parts + Text | **0.806** |

**Table 2.** Results of character grounding (**left**) and text re-identification (**right**) on the validation set of M-VAD Names [30]. Attr, Text, Visual and Grounding means joint training with the attribute module, Text Re-Id Graph, Video Re-Id Graph and Character Grounding Graph, respectively.

| Character Grounding | Accuracy |
|---|---|
| M-VAD Names [30] | 0.621 |
| MPII-MD Co-ref [33] | 0.622 |
| CiSIN w/o motion | 0.606 |
| CiSIN w/o Attr & Text | 0.651 |
| CiSIN w/o Text | 0.673 |
| CiSIN | **0.684** |

| Text Re-Id | Accuracy |
|---|---|
| BERT baseline [6] | 0.734 |
| JSFusion [48] with BERT | 0.744 |
| CiSIN w/o Grounding & Visual | 0.743 |
| CiSIN w/o Visual | 0.754 |
| CiSIN w/o Text | 0.737 |
| CiSIN | **0.767** |

fair comparison, we choose the same ResNet-50 as the CNN backbone. We also test other visual cue-based baselines [50,16,15] using the same face and body features with CiSIN. We then fine-tune the model with M-VAD Names dataset.

Despite its competence in person re-identification research, the strong-ReID-baseline significantly underperforms our model by 18.9% of accuracy drop. M-VAD Names dataset contains much more diverse person appearances in terms of postures, sizes and ill-posedness (*e.g.* extremely wide shots, partial and back views), which makes the existing re-identification model hard to attain competitive scores. In ablation studies, our model with only face embedding achieves a considerable gain of 16.6%, which implies that utilizing facial information is crucial for re-identification in movies. However, adding simple average-pooled human bbox embedding (+ Human bbox in Table 1) does not improve the performance. Instead, combining with our body-part embedding ($\mathbf{p}_m$) yields better accuracy, meaning that it can convey full-body information better than simple bbox embedding especially when human appearances are highly diverse. The variant (+ Text) in Table 1 uses the grounding and text matching to further improve video re-identification accuracy, as done in Eq.(11). We update the Video Re-Id Graph as $\mathbf{V} = \mathbf{V} + \lambda \mathbf{G}^T \mathbf{L} \mathbf{G}$, where $\lambda$ is a trainable parameter, $\mathbf{G}$ is the bipartite Grounding Graph, and $L$ is the Text Re-Id Graph. The result hints that contextual information inside the text helps improve re-identification in videos.

Table 2 presents the results of character grounding and text re-identification in M-VAD Names dataset. For character grounding, we report the results of two

**Table 3.** Quantitative results of the "Fill-in the Characters in Description" task for blinded test dataset in LSMDC 2019 challenge. * denotes the scores from the final official scores reported in LSMDC 2019 challenge slides.

| Fill-in the Characters | Accuracy |
|---|---|
| Human median* | 0.870 |
| Human (w/o video) median* | 0.700 |
| Official baseline* | 0.639 |
| YASA* | 0.648 |
| CiSIN (Visual-only) | 0.620 |
| CiSIN (Text-only) | 0.639 |
| CiSIN (Separate training) | 0.653 |
| CiSIN | **0.673** |

state-of-the-art models and different variants of our CiSIN model. Originally, the MPII-MD Co-ref [34] is designed to link the given character's name and gender to corresponding visual tracks. For fair comparison, we use the supervised version in [33] that utilizes ground truth track information, but do not provide exact character name nor gender information at test time. Our model outperforms existing models with large margins. Moreover, joint training enhances our grounding performance by 3.3%p than naïvely trained CiSIN w/o Attr & Text.

For the text-re identification task, we use JSFusion [48] and the BERT of our model with no other component as the baselines. JSFusion is the model that won the LSMDC 2017 and we modify its Fill-in-the blank model to be applicable to the new task of Fill-in the Characters. The CiSIN w/o Grounding & Visual indicates this BERT variant with additional joint training with the attribute loss. It enhances accuracy by 0.9%p and additional training with the Character Grounding graph further improves 1.1%p. Also, we report the score of using only Video Re-Id Graph and bipartite Grounding Graph, which is 73.7%. As expected, jointly training of the whole model increases the accuracy to 76.7%.

**LSMDC 2019.** Table 3 shows the results of the Fill-in the Characters task in LSMDC 2019 challenge. The baseline and human performance are referred to LSMDC 2019 official website[5]. Our model trained with both textual and visual context achieves 0.673, which is the best score for the benchmark. For the variant of CiSIN (Separate training), each component is separately trained with its own loss function. This model without joint training shows the performance drop by 2%p. Obviously, our model that lacks the text Re-Id graph significantly underperforms, meaning the association with text is critical for the task.

### 4.3 Qualitative Results

**Fill-in the Characters.** Figure 4 illustrates the results of CiSIN model for the Fill-in the Characters task with correct (left) and near-miss (right) examples. Figure 4(a) is a correct example where our model identifies characters in conversation. Most frames expose only the upper body of characters, where clothing

[5] `https://sites.google.com/site/describingmovies/lsmdc-2019`.

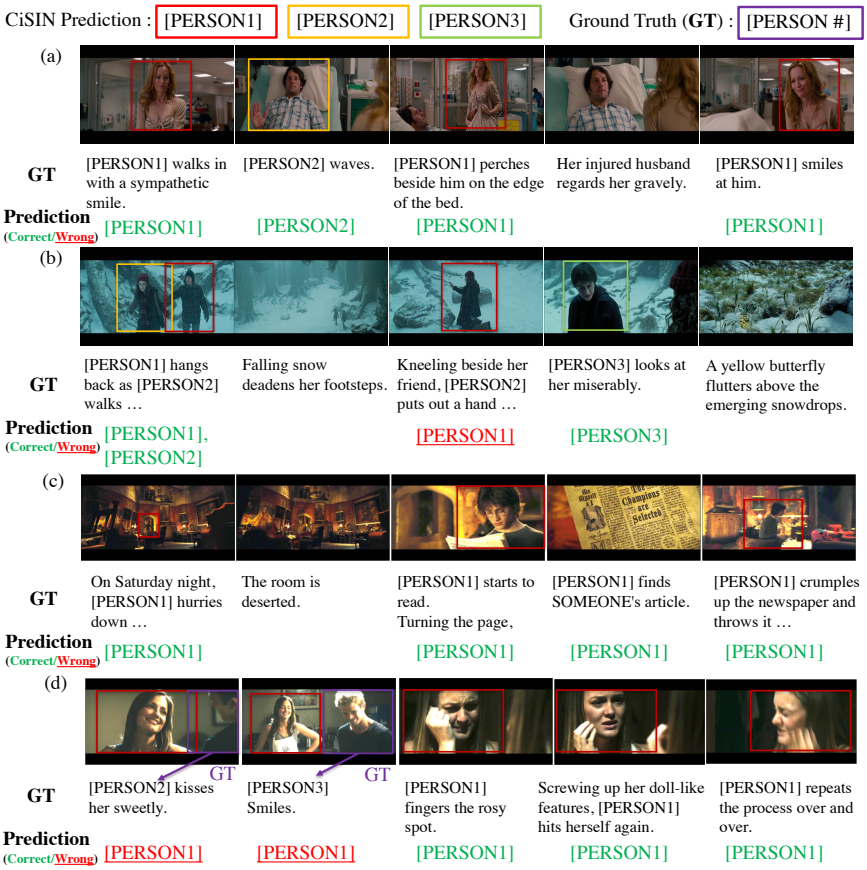


**Fig. 4.** Qualitative results of our CiSIN model. (a)-(d) are the Fill-in the Characters examples with median frames of the character grounding trajectories. Green examples indicate correct inferences by our model, while red ones with underlines in (b) and (d) are incorrect. Note that the bounding boxes are generated by our model.

and facial features become decisive clues. Figure 4(b) shows a failure case to distinguish two main characters `[PERSON1,2]` in black coats since the characters are seen in a distant side view. Figure 4(c) is another successful case; in the first and fourth video clips, the Video Re-Id Graph alone hardly identifies the character because he is too small (in the 1st clip) or only small parts (hands) appear (in the 4th clip). Nonetheless, our model can distinguish character identity from the story descriptions in text. In Figure 4(d), although the model can visually identify the same characters in neighboring video clips, the character grounding module repeatedly pays false attention to the other character. In such cases, our model often selects the candidate that stands out the most.

**Character Grounding.** Figure 5 shows character grounding results with top six visual tracks. Figure 5(a) depicts three character grounding examples

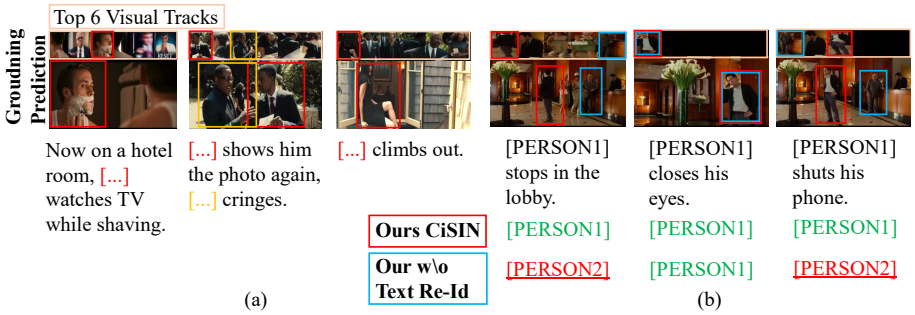


**Fig. 5.** Qualitative results of character grounding by the CiSIN model. (a) Examples of character grounding where [...] denotes a `[SOMEONE]` token. (b) Examples of Fill-in the Characters with and without the Text Re-Id Graph.

where our model uses action (*e.g.* shaving, showing a photo, climbing) and facial (*e.g.* shaving cream, cringe) information to pick the right character. Figure 5(b) shows the effect of the Text Re-Id Graph. In the first clip, two people are standing still in front of the lobby, and our grounding module would select `[PERSON2]` as an incorrect prediction without the Text Re-Id Graph. However, it can later capture the coherence of the text that the mentioned person is the same.

## 5 Conclusion

We proposed the Character-in-Story Identification Network (CiSIN) model for character grounding and re-identification in a sequence of videos and descriptive sentences. The two key representations of the model, Visual Track Embedding and Textualized Character Embedding, are easily adaptable in many video retrieval tasks, including person retrieval with free-formed language queries. We demonstrated that our method significantly improved the performance of video character grounding and re-identification in multiple clips; our method achieved the best performance in a challenge track of LSMDC 2019 and outperformed existing baselines for both tasks on M-VAD Names dataset.

Moving forward, there may be some interesting future works to expand the applicability of the CiSIN model. First, we can explore human retrieval and re-identification tasks in other domains of videos. Second, as this work only improved the re-identification of those mentioned in descriptions, we can integrate with a language generation or summarization module to better understand the details of a specific target person in the storyline.

**Acknowledgement** We thank SNUVL lab members for helpful comments. This research was supported by Seoul National University, Brain Research Program by National Research Foundation of Korea (NRF) (2017M3C7A1047860), and AIR Lab (AI Research Lab) in Hyundai Motor Company through HMC-SNU AI Consortium Fund. Gunhee Kim is the corresponding author.

## References


1. Ahmed, E., Jones, M., Marks, T.: An Improved Deep Learning Architecture for Person Re-Identification. In: CVPR (2015)
2. Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Finding Actors and Actions in Movies. In: ICCV (2013)
3. Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: CVPR (2017)
4. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person Re-Identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. In: CVPR (2016)
5. Deng, J., Guo, J., Niannan, X., Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: CVPR (2019)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT (2019)
7. Everingham, M., Sivic, J., Zisserman, A.: Hello! My name is... Buffy"–Automatic Naming of Characters in TV Video. In: BMVC (2006)
8. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR (2010)
9. Gheissari, N., Sebastian, T.B., Hartley, R.: Person Reidentification Using Spatiotemporal Appearance. In: CVPR (2006)
10. Gray, D., Tao, H.: Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In: ECCV (2008)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR (2016)
12. Hodosh, M., Young, P., Hockenmaier, J.: Framing Image Description as a Ranking Task: Data, models and Evaluation Metrics. JAIR (2013)
13. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural Language Object Retrieval. In: CVPR (2016)
14. Huang, Q., Liu, W., Lin, D.: Person Search in Videos with One Portrait Through Visual and Temporal Links. In: ECCV (2018)
15. Huang, Q., Xiong, Y., Lin, D.: Unifying Identification and Context Learning for Person Recognition. In: CVPR (2018)
16. Joon Oh, S., Benenson, R., Fritz, M., Schiele, B.: Person recognition in personal photo collections. In: ICCV (2015)
17. Kingma, D., Ba, J.: Adam: A Method for Stochastic Optimization. In: ICLR (2015)
18. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. TACL (2014)
19. Kottur, S., Moura, J.M., Parikh, D., Batra, D., Rohrbach, M.: Visual Coreference Resolution in Visual Dialog using Neural Module Networks. In: ECCV (2018)
20. Li, S., Bak, S., Carr, P., Wang, X.: Diversity Regularized Spatiotemporal Attention for Video-Based Person Re-Identification. In: CVPR (2018)
21. Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-Aware Textual-Visual Matching With Latent Co-Attention. In: ICCV (2017)
22. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person Search with Natural Language Description. In: CVPR (2017)
23. Li, W., Zhao, R.R., Xiao, T., Wang, X.: DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In: CVPR (2014)
24. Lin, D., Fidler, S., Kong, C., Urtasun, R.: Visual Semantic Search: Retrieving Videos via Complex Textual Queries. In: CVPR (2014)

25. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In: CVPR Workshop (2019)
26. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and Comprehension of Unambiguous Object Descriptions. In: CVPR (2016)
27. Nagrani, A., Zisserman, A.: From Benedict Cumberbatch to Sherlock Holmes: Character Identification in TV series without a Script. In: BMVC (2017)
28. Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., Yokoya, N.: Learning Joint Representations of Videos and Sentences with Web Image Search. In: ECCV (2016)
29. Parkhi, O.M., Rahtu, E., Zisserman, A.: It's in the Bag: Stronger Supervision for Automated Face Labelling. In: ICCV Workshop (2015)
30. Pini, S., Cornia, M., Bolelli, F., Baraldi, L., Cucchiara, R.: M-VAD Names: a Dataset for Video Captioning with Naming. MTA (2019)
31. Qi, P., Dozat, T., Zhang, Y., D. Manning, C.: Universal Dependency Parsing from Scratch. In: CoNLL 2018 UD Shared Task (2018)
32. Ramanathan, V., Joulin, A., Liang, P., Fei-Fei, L.: Linking People in Videos with "their" Names Using Coreference Resolution. In: ECCV (2014)
33. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of Textual Phrases in Images by Reconstruction. In: ECCV (2016)
34. Rohrbach, A., Rohrbach, M., Tang, S., Oh, S.J., Schiele, B.: Generating Descriptions with Grounded and Co-Referenced People. In: CVPR (2017)
35. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie Description. IJCV (2017)
36. Seo, P.H., Lehrmann, A., Han, B., Sigal, L.: Visual Reference Resolution using Attention Memory for Visual Dialog. In: NIPS (2017)
37. Shen, Y., Lin, W., Yan, J., Xu, M., Wu, J., Wang, J.: Person Re-Identification with Correspondence Structure Learning. In: ICCV (2015)
38. Sivic, J., Everingham, M., Zisserman, A.: "Who are you?"-Learning Person Specific Classifiers from Video. In: CVPR (2009)
39. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven Deep Convolutional Model for Person Re-Identification. In: ICCV (2017)
40. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: "Knock! Knock! Who is it?" Probabilistic Person Identification in TV-series. In: 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. Torabi, A., Tandon, N., Sigal, L.: Learning Language-Visual Embedding for Movie Understanding with Natural-Language. arXiv:1609.08124 (2016)
43. Vendrov, I., Kiros, R., Fidler, S., Urtasun, R.: Order-embeddings of Images and Language. In: ICLR (2016)
44. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval. In: ACM MM (2017)
45. Wojke, N., Bewley, A., Paulus, D.: Simple Online and Realtime Tracking with a Deep Association Metric. In: ICIP (2017)
46. Xu, R., Xiong, C., Chen, W., Corso, J.J.: Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework. In: AAAI (2015)
47. Yan, Y., Zhang, Q., Ni, B., Zhang, W., Xu, M., Yang, X.: Learning Context Graph for Person Search. In: CVPR (2019)
48. Yu, Y., Kim, J., Kim, G.: A joint sequence fusion model for video question answering and retrieval. In: ECCV (2018)

49. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint Face Detection and Alignment using Multitask Cascaded Convolutional Networks. IEEE Signal Proc (2016)
50. Zhang, N., Paluri, M., Taigman, Y., Fergus, R., Bourdev, L.: Beyond frontal faces: Improving person recognition using multiple cues. In: CVPR (2015)
51. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable Person Re-Identification: A Benchmark. In: ICCV (2015)
52. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: MARS: A Video Benchmark for Large-Scale Person Re-Identification. In: ECCV (2016)
53. Zhou, X., Wang, D., Krähenbühl, P.: Objects as Points. arXiv:1904.07850 (2019)