Aggregate Tracklet Appearance Features for Multi-Object Tracking

Long Chen, Haizhou Ai, *Senior Member, IEEE*, Rui Chen, and Zijie Zhuang

***Abstract*—Multi-object tracking (MOT) has wide applications in the fields of video analysis and signal processing. A major challenge in MOT is how to associate the noisy detections into long and continuous trajectories. In this letter, we address the association problem at the tracklet-level, and mainly focus on the appearance representation designed for tracklets. A multitask convolutional neural network is proposed to learn the discriminative features and spatial-temporal attentions jointly. In particular, we decompose an object in a static image with spatial attentions, and then aggregate multiple features in a tracklet based on the temporal attentions. Appearance misalignment that caused by occlusion and inaccurate bounding is then mitigated by multi-feature aggregation. Experimental results on two challenging MOT benchmarks have demonstrated the effectiveness of the proposed method and shown significant improvement on the quality of tracking identities.**

***Index Terms*—Multi-object tracking, tracklet association, appearance model, spatial-temporal attention.**

I. Introduction

MULTI-OBJECT Tracking (MOT) in a monocular camera plays a crucial role in the applications of video analysis and signal processing. The goal of this task is to identify targets of interest automatically and estimate their trajectories across frames. Benefiting from recent advances in object detection, *tracking-by-detection* dominates the field of multi-object tracking. The main idea is to first apply an object detector for each frame, and then associate detection responses across frames into trajectories. How to link these detections, namely data association, is the major problem under such a tracking framework.

Existing data association approaches can be divided into two categories, from the *online mode* where only the current and previous frames are considered [1]–[4], to the *offline mode* that utilizing global information from future frames [5]–[8]. In the online mode, detections are linked with existing trajectories frame-by-frame, which is straightforward and lightweight but often suffers from the uncertainties of the observation [1]. The ambiguity caused by target clustering, occlusion and missing detections is hard to be eliminated with limited observation from a single frame. Offline methods, in contrast to online tracking, are designed for robust long trajectories by exploiting the global observations from the entire video. In order to reduce the computational complexity, it is a common practice for the offline method to perform the association at the tracklet-level [7], [9]–[12]. They link detections to tracklets (short track fragments) with online tracking first, and then associate the tracklets into long and continuous trajectories.

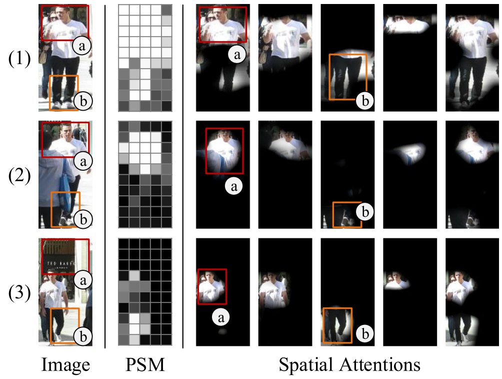

Fig. 1. Image patches along with the spatial attentions. The regions (a) and (b) on the image are not matched accross frames due to occlusion and inaccurate bounding. With the guidance from spatial attentions, (a) and (b) learn to focus on the corresponding positions of the object.

In that case, the basis of data association algorithm is the affinity measure between detections or tracklets. Many works were done with a carefully-designed affinity measure. Wang *et al.* [9] calculate the affinity between tracklets based on the color histogram and HOG feature [13]. Xiang *et al.* [14] employ an appearance model trained with convolutional neural network (CNN) to measure the affinity between detections. With significant progress in deep learning, exploiting CNN-based appearance features is a trend in this area [3], [14]–[19]. The discriminative feature learned from large-scale dataset [20] offers superior association performance [2], [16], [21]. These methods mainly focus on the static images. When switching to the tracklet association, appearance misalignment between frames is however less investigated. As illustrated in the first column of Fig. 1, noisy detections such as occlusion and inaccurate bounding might result in content mismatch at the same location of different images, though these images are cropped from the same target.

Manuscript received July 30, 2019; revised September 3, 2019; accepted September 3, 2019. Date of publication September 11, 2019; date of current version September 25, 2019. This work was supported by the Natural Science Foundation of China under Project 61521002. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Marco F. Duarte. *(Corresponding author: Long Chen.)*

The authors are with the Computer Science and Technology Department, Tsinghua University, Beijing 100084, China (e-mail: l-chen16@mails.tsinghua.edu.cn; ahz@mail.tsinghua.edu.cn; chenr18@mails.tsinghua.edu.cn; zhuangzj15@mails.tsinghua.edu.cn).

Digital Object Identifier 10.1109/LSP.2019.2940922

In this letter, we tackle the problem of tracklet association for multi-object tracking, and specifically focus on the appearance model designed for tracklets. Ideally, tracklet can provide more information than a static image that used in previous appearance models. To take advantage of it, a multitask CNN model is proposed to learn the discriminative features and spatial-temporal attentions jointly. We decompose an object in a static image with spatial attentions, and then aggregate features of images within the tracklet based on the temporal attentions. The motivation behind it is to obtain a robust affinity measure by mitigating the appearance misalignment problem with multi-feature aggregation. Our contributions can be summarized as follows: 1) A novel spatial-temporal attention mechanism is proposed for MOT, without introducing additional expensive annotations. 2) Appearance misalignment problem is addressed by aggregating tracklet features with the spatial-temporal attentions. 3) Extensive experiments are conducted on two widely used MOT datasets to demonstrate the effectiveness of the proposed method.

II. Tracklet Appearance Model

In this section, we present an appearance model with spatial-temporal attentions that designed for tracklet association. We first introduce our attention mechanism that inspired from position-sensitive mask [22], the proposed network architecture and training method are detailed afterwards.

A. Position-Sensitive Mask

Part-aligned representation with spatial attentions has been proposed to address the appearance misalignment in the field of MOT [16] and person re-identification (ReID) [23], [24]. Our method is different from previous methods, which require additional expensive supervision from pose estimation or semantic segmentation. In contrast, inspiring from position-sensitive mask (PSM), we train the attention model with only supervision from the foreground/background classification, of which the annotations are within reach in MOT.

Given an image patch I, the training objective of PSM is to classify the image into foreground and background with a predicted probability $p_{cls}(y|I)$. Instead of obtaining the probability directly from a regression layer, PSM is a set of feature maps $\mathbb{Z} = \{\mathbf{z}_k\}$, where each feature map has a spatial dimension of $w \times h$. Setting the number of feature maps $|\mathbb{Z}|$ to $w \times h$, the foreground probability is then defined as

$$p_{cls}(y|I) = \sigma\left(\frac{1}{w \times h}\sum_{i=1}^{h}\sum_{j=1}^{w}\mathbf{z}_k(i,j)|_{k=(i-1)\times w+j}\right), \quad (1)$$

where $\sigma(\cdot)$ represents the sigmoid function, and $\mathbf{z}_k(i,j)$ is the response in the $(i,j)_{th}$ spatial position for the k_{th} feature map. In this way, each feature map responses to a specified object position that explicitly encoded by the classification criterion. For example, as illustrated in Fig. 2(a), with $w \times h = 3 \times 3$, we have 9 feature maps respond to the *top-left*, *top-center*, *top-right*,..., *bottom-right* of the object, respectively. These position-sensitive feature maps are utilized to form the spatial-temporal attention, as detailed in the following section.

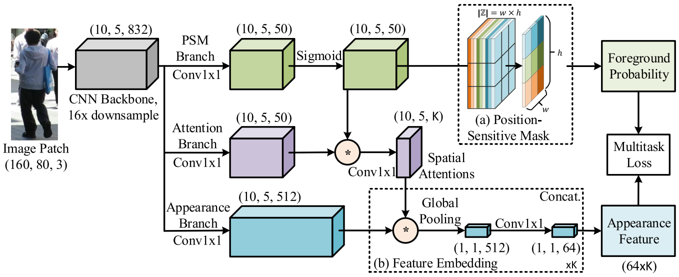

Fig. 2. Architecture of the appearance model. The numbers in the figure represent the dimensions of output features. (a) Illustartion of position-sensitive mask. (b) Feature embedding with global average pooling.

B. Appearance Model With Spatial-Temporal Attention

As shown in Fig. 2, the proposed appearance model consists of a CNN backbone and three sub-branches following to the backbone: *PSM branch*, *Attention branch*, and *Appearance branch*. All the layers and corresponding feature dimensions of the branches are detailed in the figure.

For an image patch I from the tracklet $T = \{I_1, \ldots, I_N\}$, feature extraction is performed in three steps: (1) Position-sensitive feature maps $\mathbb{Z}$ and foreground probability $p_{cls}(y|I)$ are first obtained from the PSM branch. (2) We apply $\mathbb{Z}$ on the attention branch with an element-wise multiplication. K spatial attention maps are then obtained after a 1×1 convolutional layer. (3) On the appearance branch, we apply the spatial attentions to the discriminative features and extract a 64-d feature vector for each attention map. The feature vector $f_{img}(I)$ of the image is finally concatenated from the K feature vectors.

Fig. 1 shows some examples about spatial attentions. In that case, each attention map can be considered as a linear combination over the responses of different object positions. The weights of the combination are formulated and trained by the convolutional layer. With the part-aligned image-level features, we employed the foreground probability $p_{cls}(y|I)$ as the temporal attention to aggregate multiple features:

$$f_{trk}(T) = \frac{\sum_{i=1}^{N} f_{img}(I_i) * p_{cls}(y|I_i)}{\sum_{i=1}^{N} p_{cls}(y|I_i)}. \quad (2)$$

where I_i is the i_{th} image patch in the tracklet. The motivation of the whole model is to mitigate the appearance misalignment problem when coping with the spatial-temporal attentions.

C. Multitask Learning

As mentioned above, the model requires only the annotations of target identities, which is within reach in the MOT dataset. With the annotations, we train the model in two tasks: discriminative appearance feature learning, and image classification between foregrounds and backgrounds.

We form the training data as a set of tracklet triplets $\{\langle T_i, T_j, T_k\rangle\}$, where $\langle T_i, T_j\rangle$ is a pair of tracklets from the same target, and $\langle T_i, T_k\rangle$ is that from two different targets. We use Euclidean distance to measure the distance between features: $d_{ij} = ||f_{trk}(T_i) - f_{trk}(T_j)||$. Given N training triplets in a batch, the triplet loss [25] for the discriminative appearance feature learning is defined as:

$$L_a = \frac{1}{N} \sum_{\langle T_i, T_j, T_k\rangle} [d_{ij} - d_{ik} + m]_+, \tag{3}$$

where $[x]_+$ represents $\max(0, x)$, and $m = 0.2$ is a predefined margin. The training goal of this loss function is to decrease the distance between features from the same target and increase that from different targets.

As for the classification task, image patches around the targets are employed as training data. The target patches are regarded as positive samples and we randomly replace 30% of the tracklets with background patches as negative samples. Binary cross entropy are utilized for the classification loss:

$$L_c = \frac{1}{N} \sum_{i=1}^{N} y_i^* log(y_i) + (1 - y_i^*) log(1 - y_i), \tag{4}$$

where $y_i = p(y|I_i)$ and $y^* \in \{0, 1\}$ is the ground truth label of the image patch. Note that negative patches are ignored when calculating the appearance loss. The final multitask loss is then defined as:

$$L = L_a + \lambda_c L_c, \tag{5}$$

where λ_c is the scale factor to balance the appearance and classification loss.

III. Multi-Object Tracking Framework

A. Affinity Measure for Tracklets

Affinity measure is the basis of any data association algorithm. Given two tracklets $\langle T_i, T_j\rangle$, we define the appearance affinity as

$$A_a(T_i, T_j) = \lambda_a - ||f_{trk}(T_i) - f_{trk}(T_j)||, \tag{6}$$

where λ_a is a predefined threshold. In this letter, we mainly focus on the tracklet appearance model, therefore, we only add a simple constraint on location and motion. A delicate motion model might complement the appearance affinity, however, it is out of the scope of this letter. The constraint is defined as

$$A_m(T_i, T_j) = \mathbb{1}(\triangle t_{ij} > 0) \cdot \mathbb{1}(\triangle s_{ij} / \triangle t_{ij} < \lambda_v), \tag{7}$$

where $\mathbb{1}$ is the indicator function, $\triangle t_{ij}$ is the time interval of two tracklets, $\triangle s_{ij}$ is the spatial distance between the end location of T_i and the start location of T_j. In that case, λ_v can be considered as the threshold of moving speed. This simple constraint ensure that tracklets overlapping in time or far away in space will not be associated. We set $\lambda_a = 0.33$ and $\lambda_v = 1.2 \times v_{max}$ where v_{max} is the maximum speed among all tracklets in the time window.

B. Tracklet Association

We formulate tracklet association in the time window as a graph decomposition problem. The graph is defined as a weighted undirected graph $G = (V, E, W)$ on $V = \{T_i\}$. The weight w_{ij} of an edge $\langle T_i, T_j\rangle \in E$ is calculated by fusing the appearance affinity and motion constraint:

$$w_{ij} = A_a(T_i, T_j) \cdot A_m(T_i, T_j). \tag{8}$$

Finally, we solve the graph decomposition problem by maximizing the total affinity with binary integer programming [26]:

$$\mathbf{x}^* = \underset{X}{\text{argmax}} \sum_{\langle T_i, T_j\rangle \in E} w_{ij} x_{ij}, \tag{9}$$

subject to

$$x_{ij} \in \{0, 1\} \quad \forall \langle T_i, T_j\rangle \in E, \tag{10}$$

$$x_{ij} + x_{jk} \leq 1 + x_{ik} \quad \forall \langle T_i, T_j\rangle, \langle T_i, T_k\rangle, \langle T_j, T_k\rangle \in E, \tag{11}$$

where X is the set of all possible combinations of assignments to the binary variables x_{ij}. Equation (11) enforces the transitivity of the association problem, that is, $\langle T_i, T_k\rangle$ must be connected in the graph if $\langle T_i, T_j\rangle$ and $\langle T_j, T_k\rangle$ are connected. We associate the tracklets in the time window into long and continuous trajectories based on the optimal solution $\mathbf{x}^*$.

IV. Experiments

A. Implementation

Experiments are conducted on two widely used multiple people tracking benchmark: MOT16 and MOT17 [28]. Since the proposed method does not rely on human specified attributes, we argue that the method is equally applicable to other kinds of objects, e.g., vehicle [19].

Specifically, we utilize the online tracker MOTDT [2] as the baseline method to generate tracklets. As for the appearance model, GoogLeNet [29] is utilized as the backbone, and the network is trained for 10k steps using Adam optimizer [30], with the learning rate of 1e-4 and the scale factor λ_c of 3. We set the number of attention maps K to 8 and the size of time window to 70 based on the experiments detailed in the Section IV-D. Note that the tracking performance is heavily related to the quality of detection, for a fair comparison, we only use the public detection provided by the benchmark in the following experiments.

B. Evaluation Metrics

Following existing methods [2], [12], [16], we evaluate the proposed method on multiple metrics, including multi-object tracking accuracy (MOTA) [31], ID recall (IDR), ID precision (IDP) and ID F1 score (IDF1) [32], mostly tracked targets (MT), mostly lost targets (ML) [33], numbers of ID switches (IDS), as well as tracking speed (FPS). MOTA reflects the overall performance since it considers multiple aspects including precision, recall and identity switches. As a complement, IDF1 is specifically designed to measure the identity precision and

TABLE I
COMPARISON ON MOT16 AND MOT17 TEST SETS. BEST IN BOLD, SECOND BEST IN BLUE

Benchmark	Method	MOTA(%)↑	IDF1(%)↑	IDP(%)↑	IDR(%)↑	MT(%)↑	ML(%)↓	IDS↓	FPS↑
MOT16	STAM16 [3]	46.0	50.0	71.5	38.5	14.6	43.6	473	0.2
	NOMT [11]	46.4	53.3	73.2	41.9	**18.3**	41.4	**359**	2.6
	MOTDT [2] (baseline)	47.6	50.9	69.2	40.3	15.2	38.3	792	**20.6**
	LMP [16]	48.8	51.3	71.1	40.1	18.2	40.1	481	0.5
	TNT [10]	49.2	**56.1**	**75.9**	**44.5**	17.3	40.3	606	0.7
	HCC [21]	49.3	50.7	71.1	39.4	17.8	39.9	391	0.8
	NOTA (proposed)	**49.8**	55.3	75.3	43.7	17.9	**37.7**	614	19.2
MOT17	MTDF17 [27]	49.6	45.2	58.1	37.0	18.9	**33.1**	5567	1.2
	MOTDT [2] (baseline)	50.9	52.7	70.4	42.1	17.5	35.7	2474	**18.3**
	JBNOT [12]	**52.6**	50.8	64.8	41.7	**19.7**	35.8	3050	5.4
	NOTA (proposed)	51.3	**54.5**	**73.5**	**43.2**	17.1	35.4	**2285**	17.8

TABLE II
RESULTS WHEN USING DIFFERENT APPEARANCE MODELS

Appearance Model	IDF1(%)	MOTA(%)	IDS
HOG [13]	46.7	41.8	209
Color Histogram	53.4	42.4	162
Color Hist. + Segmentation	51.9	41.9	244
Image CNN [2]	54.7	42.7	155
Image CNN + Segmentation	55.1	42.9	160
Video-based ReID [35]	56.8	43.0	149
Our Method	**58.4**	**43.3**	**106**

recall. Considering that the appearance model mainly affects the identity quality, we adopt both MOTA and IDF1 as our primary metrics in the following analysis.

C. Comparison on MOT Benchmark

The comparison on MOT16 and MOT17 test sets is shown in Table I. The proposed multi-feature aggregation and tracklet association improves the evaluation metrics significantly comparing to the baseline method [2], which contains an appearance model on static images. While LMP [16] introduces extra information from pose estimation for the appearance learning, our method, with limited data, still achieves better performance in terms of both MOTA and IDF1. Moreover, TNT [10] is a state-of-the-art tracklet association framework that combines geometric information, temporal-spatial dependency, and face recognition features [34]. Benefiting from the proposed appearance model, our method achieves competitive identity quality while is much more efficient compared to TNT. On MOT17 benchmark, JBNOT [12] achieves the best MOTA by using additional body and joint detections while ignores the appearance features. Our method outperforms JBNOT by 8.7% IDP and 3.7% IDF1, indicating the effectiveness of the proposed appearance model.

D. Analysis on Validation Set

We further investigate the performance when using different appearance models on MOT16 training set. Each sequences in the original training set is split by a ratio of 3:2 to form the training and validation sets for the following analysis.

The experimental results are presented in Table II. All the hyper-parameters, except the affinity threshold λ_a, are fixed among different models. We choose λ_a with a grid search, selecting the value with the best IDF1 score. Four different appearance models are considered: HOG feature [13], color histogram, image-based CNN model [2], as well as a video-based person ReID method that contains the temporal attention [35]. The proposed method achieves the best performance on IDF1, MOTA and IDS. Moreover, appearance model with semantic segmentation is also investigated. We generate object masks using Deeplab [36] trained on VOC dataset [37]. The masks are utilized to filter out background pixels for color histogram, or served as an extra input channel for the CNN model. As shown in the table, adding segmentation has little effect on the CNN model and even spoils the color histogram. That is because, semantic segmentation can not handle the case of intra-category occlusion, which is a major reason of the appearance misalignment problem in MOT.

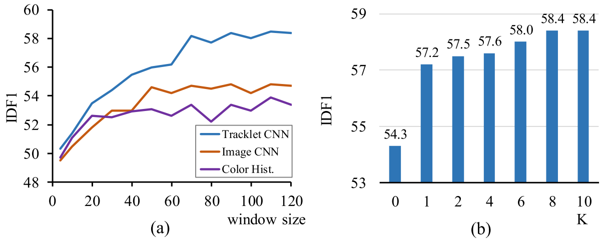

Fig. 3. The performance under different settings of (a) window size and (b) the number of attention maps. The appearance model without attention branch is represented by $K = 0$.

The performance under different settings of window size and number of attention maps K is presented in Fig. 3. The length of tracklet for the appearance model is affected by the window size. The performance of color histogram is saturated when the window size is increased beyond 20. This indicates that, increasing the length of tracklets introduces more noises from backgrounds than useful features, resulting in ambiguities in the association. In contrast, our model benefits from large window size, demonstrating the effectiveness of the proposed multi-feature aggregation method.

V. CONCLUSION

In this letter, we present a novel tracklet appearance model when coping with spatial-temporal attentions for multi-object tracking. The model blends tracklet appearance and position-sensitive mask in a multitask convolutional neural network, where the supervision from foreground/background classification guides the attention learning. With the proposed affinity measure, tracklet association is formulated as a graph decomposition probelm and solved in binary integer programming. Experimental results have shown the superior performance of the proposed method comparing to the state-of-the-art methods on MOT16 and MOT17 benchmarks.

References

[1] F. Solera, S. Calderara, and R. Cucchiara, "Learning to divide and conquer for online multi-target tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4373–4381.

[2] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2018, pp. 1–6.

[3] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4846–4855.

[4] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang, "Online multi-object tracking with dual matching attention networks," in *Proc. Euro. Conf. Comput. Vis.*, 2018, pp. 379–396.

[5] E. Levinkov *et al.*, "Joint graph decomposition amp; node labeling: Problem, algorithms, applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1904–1912.

[6] M. Keuper, S. Tang, B. Andres, T. Brox, and B. Schiele, "Motion segmentation multiple object tracking by correlation co-clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, pp. 1–1, Art. no. 8493320, doi: 10.1109/TPAMI.2018.2876253.

[7] B. Wang, G. Wang, K. L. Chan, and L. Wang, "Tracklet association with online target-specific metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1234–1241.

[8] C. Wang, H. Liu, and Y. Gao, "Scene-adaptive hierarchical data association for multiple objects tracking," *IEEE Signal Process. Lett.*, vol. 21, no. 6, pp. 697–701, Jun. 2014.

[9] J. Wang, Y. Guo, X. Tang, Q. Hu, and W. An, "Semi-online multiple object tracking using graphical tracklet association," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1725–1729, Nov. 2018.

[10] G. Wang, Y. Wang, H. Zhang, R. Gu, and J.-N. Hwang, "Exploit the connectivity: Multi-object tracking with trackletnet," 2018, *arXiv:1811.07258*.

[11] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3029–3037.

[12] R. Henschel, Y. Zou, and B. Rosenhahn, "Multiple people tracking using body and joint detections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, pp. 886–893.

[14] J. Xiang, G. Zhang, J. Hou, N. Sang, and R. Huang, "Multiple target tracking by learning feature representation and distance metric jointly," in *Proc. British Mach. Vis. Conf.*, 2018, p. 139. [Online]. Available: https://dblp.org/rec/bibtex/conf/bmvc/XiangZS0H18

[15] L. Chen, H. Ai, C. Shang, Z. Zhuang, and B. Bai, "Online multi-object tracking with convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 645–649.

[16] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3701–3710.

[17] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 3645–3649.

[18] C. Kim, F. Li, and J. M. Rehg, "Multi-object tracking with neural gating using bilinear LSTM," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 208–224. [Online]. Available: https://dblp.org/rec/bibtex/conf/eccv/KimLR18

[19] Z. Tang, G. Wang, H. Xiao, A. Zheng, and J. Hwang, "Single-camera and inter-camera vehicle tracking and 3D speed estimation based on fusion of visual and semantic features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 108–1087.

[20] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.

[21] L. Ma, S. Tang, M. J. Black, and L. V. Gool, "Customized multi-person tracker," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 612–628. [Online]. Available: https://dblp.org/rec/bibtex/conf/accv/MaTBG18

[22] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[23] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3239–3248.

[24] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[25] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.

[26] E. Ristani and C. Tomasi, "Tracking multiple people online and in real time," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 444–459. [Online]. Available: https://dblp.org/rec/bibtex/conf/accv/RistaniT14

[27] Z. Fu, F. Angelini, J. Chambers, and S. M. Naqvi, "Multi-level cooperative fusion of GM-PHD filters for online multiple human tracking," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2277–2291, Sep. 2019.

[28] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," 2016, *arXiv:1603.00831*.

[29] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015. [Online]. Available: https://dblp.org/rec/html/journals/corr/KingmaB14

[31] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP J. Image Video Process.*, vol. 2008, 2008, Art. no. 246309. [Online]. Available: https://dblp.org/rec/bibtex/journals/ejivp/BernardinS08

[32] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Euro. Conf. Comput. Vis. Workshops*, 2016, pp. 17–35. [Online]. Available: https://dblp.org/rec/bibtex/conf/eccv/RistaniSZCT16

[33] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2953–2960.

[34] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[35] J. Gao and R. Nevatia, "Revisiting temporal modeling for video-based person reid," 2018, *arXiv:1805.02104*.

[36] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[37] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.