2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Towards Understanding and Inferring the Crowd: Guided Second Order Attention Networks and Re-identification for Multi-object Tracking

Niraj Bhujel [1], Li Jun[2], Yau Wei Yun[2] and Han Wang[1]

***Abstract*— Multi-human tracking in the crowded environment is a challenging problem due to occlusions, pose change, viewpoint variation and cluttered background. In this work, we propose a robust feature learning for tracking-by-detection methods based on second-order attention network that can capture higher-order relationships between salient features at the early stages of Convolutional Neural Network (CNN). Guided Second-Order Attention Network (GSAN) that, unlike the existing attention learning methods which are weakly-supervised, uses a supervisory signal based on the quality of the self-learned attention maps. More specifically, GSAN looks into the attended maps of a person having the highest confidence and supervise itself to look into the correct regions in the images of the person. Attention maps learned this way are spatially aligned and thus robust to camera-view changes and body pose variations. We verify the effectiveness of our approach by comparing with the state-of-the-art methods on challenging person re-identification and multi object tracking (MOT) datasets.**

I. INTRODUCTION

Autonomous robotic navigation in less-crowded or at well structured space is considered as a well-solved problem nowadays, thanks to the substantial research efforts in the last decades. Within a pre-mapped environment, modern path planning algorithms such as [1], [2] are able to drive the robot to an arbitrary goal position in the map. Although these planners enable the robots to avoid objects along their navigation path, those objects are usually assumed to be static and few planners exist. As a result, creating a collision-free path in a cluttered environment with a number of dynamic objects, especially a crowd of moving pedestrians, is still challenging task.

With the understanding of how people move and the ability to predict where they are likely to go in the following instants, new generation of autonomous system will augment its navigation behaviours in more crowded areas; avoid collision with moving people, follow their masters and escape from extreme crowded spaces. Considering this, tracking and predicting the motion of dynamic objects (e.g. pedestrian) from a robot perspective is a vital step to allow autonomous robot to operate in environments like hospitals, subway stations, and shopping malls. Such a problem can be referred as multi-object tracking (MOT), where the goal is to estimate the location of all interested objects in a video and maintain their identities consistently such that an individual trajectory can be associated to each particular object. An intuitive way to address the MOT problem is to leverage the previous research achievements for single object tracking. Those visual trackers, based on discriminative appearance features learned either by correlation filters [3] or from deep convolution neural networks (CNNs) [4], have been designed to track a pre-identified template by examining the correlative similarity between two consecutive frames. In crowded environments, appearance variability due to group interactions, occlusions, illumination changes and similar looking pedestrians make the appearance model unreliable in tracking objects frame by frames.

Recently, tracking-by-detection approach has emerged as the preferred methods. These methods simplify the multi-object tracking into two steps: first, localize the targets independently at each frames using bounding boxes (detections), and secondly, perform association between the bounding boxes over consecutive frames yielding a trajectory for all tracked identities. Usually, a similarity metric based

Fig. 1: Illustration of detection association using our proposed model. From top to bottom, bounding box detections at frame $t-1$ and correctly associated bounding box detections at frame t. The corresponding appearance S_a and motion affinity S_m values ranges between $0-1$ (0 represent lowest similarity and 1 highest similarity).

[1]Niraj Bhujel and Assoc. Prof. Han Wang are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. bhuj0001@e.ntu.edu.sg, hw@ntu.edu.sg

[2]Dr. W.Y. Yau and Dr. Li. Jun is with the Institute for Infocomm Research, A*STAR, Singapore. wyyau@i2r.a-star.edu.sg, jli@i2r.a-star.edu.sg

on the appearance and motion of the tracked target is used to make the association decision. Although this approach appears to be straightforward, the tracking performance is largely influenced by the quality of object detectors i.e. false positives, inaccurate estimation of the bounding box etc. and the ability of the association methods to handle similar looking identities, occlusions, variation in view-points, pose changes and background clutter.

Based on the fact that identities having very similar appearances have distinct attributes i.e. carry bag, tee shirt logo, watch, shoes etc. that can uniquely identify them in the crowds, we propose a Guided Second-order Attention Network (GSAN) to learn such fine-grained salient features of the pedestrians. The advantages of the proposed approach is two-fold. First, GSAN utilizes the supervisory signal to guide the attention learning process in contrast to weakly-supervised approaches [5], [6], [7]. The supervisory signal allows the attention network to determine where to look into the image regions for the salient features of the person. Second, given the fact that the earlier stages of CNN like ResNet [8] have higher spatial relationships among the features, which are usually ignored by the current attention models, the proposed approach capture this higher-order statistics via a second-order attention module. Higher-order relationships plays crucial role in distinguishing pedestrians having similar appearances by capturing subtle differences between visual parts [7].

The GSAN typically produces visual attention maps that highlights the image regions having discriminative features. An evaluation network checks the quality of the current attention maps based on the confidence scores of the predicted ground truth labels. The confidence scores indicates whether the attended features are able to lead to any improvement in the prediction accuracy. When the attentions are incorrectly aligned, i.e. decrease in confidences, the evaluation network generates a supervisory signal to guide the attention network to look into the correct regions. More specifically, the evaluator computes the squared distance between the lowest confidence attentions maps and the attention maps having the highest confidence score. The difference, when back-propagated, allows the attention network to correct it's mistakes and adapt to generate accurate attention maps. Using such strategy to train attention models, the learned attention maps are invariant to body pose and camera view as the visual parts are spatially aligned by the GSAN. We validate the effectiveness of our approach in popular benchmarks including Market-1501 and MOT challenges.

II. Related Work

A. Tracking-by-Detection

In *tracking-by-detection* framework, detections generated by a object detector at different frames are linked with the correct targets over time yielding a set of trajectories of all targets. The detections assignment is often formulated as global optimization problem and solved by graphical model where each node represent a detection and edges connect detections across frames. The aim is to find disjoint paths by minimizing cost [9] or maximizing flow [10]. While the generated paths do not overlap or merge, it require careful measures to handle overlapping detections per person which occur frequently in object detectors. The advancement in deep feature learning [8] and rapid growth in object detector's performances beyond human level have created new opportunities [11]. Traditional methods like [12] are revisited in a tracking-by-detection scenario [13] and is promising direction for multi-object tracking.

B. Appearance Models and Re-identification

Appearance model aims to extract the target specific visual features, usually color features, that are re-identified or associated at different frames maintaining a track for each target. A robust pairwise similarity cost matrix is either build or learned from visual features to solve the data association problem [13], [14], [15]. In [13], [14], near-online data association utilizing target specific feature from CNN features is proposed. The data association can be learned by directly from visual features as in [15].

The current state of the art in person re-identification relies on deeply learned features from static images [16], [17], [18]. [16] proposes camera-invariant descriptor to address image style variations caused by different cameras. [17] divide the spatial region of the feature maps into 6-parts followed by part classifier, [18] proposes negative hard sample mining with triplet loss that minimize distance between positive pairs while pushing the negative pairs away. [19] proposed to use adaptive weighted triplet loss of [18] in the context of multi-camera tracking. In [20], state-aware re-identification features like occlusion status and orientation information are utilized in Re-ID model. [21] design and train deep networks for re-identifying persons later used for associating person hypotheses. Moreover, [5], [6] exploits spatial attention and temporal attention to deal with pose variations, body-part occlusions, and misaligned detections.

C. Motion Models and Trajectory Prediction

Several works on multi-target tracking make use of motion information to gain performances improvement. The most common information is the velocity of pedestrians that is considered either constant [22] or expressed with more elaborative Social Force Model [23]. The constant velocity models assumes that all pedestrians maintain linear fix velocity independently while the latter assumes that motion of each pedestrians is influenced by the interaction among subjects and their final destination. [24] [25] exploit deep learning to learn such interactions, [24] proposed to use single LSTM per pedestrian for learning motion behaviour and the interaction among pedestrians. The model is further extended to include interaction between environment and pedestrians in [25].

III. Our Approach

A. Global Discriminative Deep Feature Learning

In the context of MOT, the visual features of pedestrians are largely influenced by occlusions, camera viewing condition and appearance characteristics of the person. To mitigate

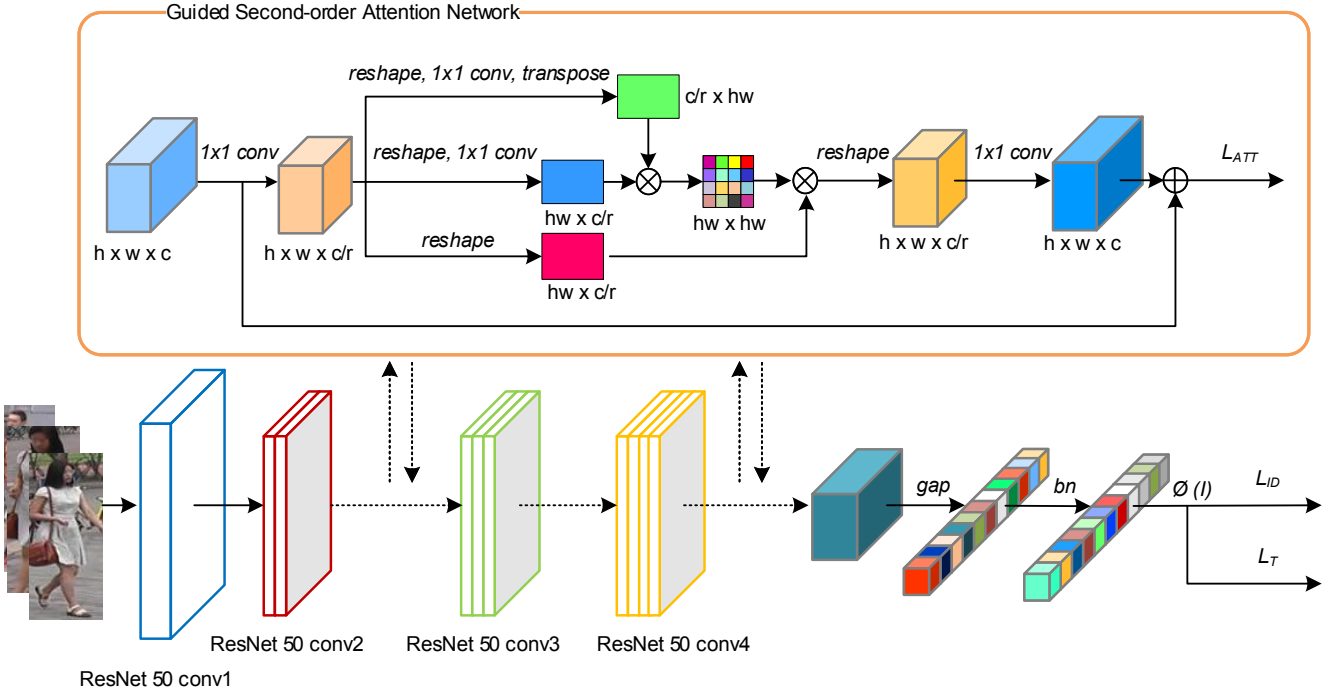

Fig. 2: An overview of our proposed approach for global and local features learning. The global branch, composed of global-average-pooling(gap), batch-normalization(bn) and classifier, serves as supervision for learning the global features $\phi(I)$. The GSAN module is responsible for identifying discriminative local features and can be placed at early stages of the ResNet50.

the adverse effects, first, we learn global discriminative features of pedestrians using identity embedding, batch-hard mining and data-augmentation. Denote the intermediate feature map of an image $\psi(I) \in \mathbb{R}^{W \times H \times C}$ are obtained from deep convolution neural network (CNN) such as ResNet50 [8]. The objective is to learn the discriminative visual feature $\phi(I) \in \mathbb{R}^{D}$ from $\psi(I)$;

$$\phi(I) = \phi(\psi(I)) = W_\phi \psi(I) + b_\phi \tag{1}$$

Given N images belonging to K persons, we employ an identity loss over $\phi(I)$.

$$L_{ID} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_{n,k} \log\left(\frac{\exp(W_k^T \phi(I_n))}{\sum_{k=1}^{K}\exp(W_k^T \phi(I_n))}\right) \tag{2}$$

where $y_{n,k} = 1$ if image I_n belongs to person k otherwise $y_{n,k} = 0$. W_k is the classifier parameter associated with the k^{th} person. The identity loss function essentially maximize the likelihood of predicting the correct identity for each training images.

The global features $\phi(I)$ is the projections of feature space of all images onto the embedding space $\mathbb{R}^D$. We want to ensure that the projected points of an image of a specific identity is closer to all other images of the same identity and further to any images of any other identity. Thus, we implement batch hard triplet loss with hard sample mining[18] as:

$$L_T = \sum_{i=1}^{K}\sum_{a=1}^{N}\left[m + \| \phi(I_a^i) - \phi(I_p^i) \|_2^2 - \| \phi(I_a^i) - \phi(I_n^i) \|_2^2\right]_+ \tag{3}$$

where $[.]_+$ indicates the function $\max(0, [.])$, I_a^i is an image (anchor) of an specific identity i, I_p^i is the images (positive) of the same identities, I_n^i is the images (negative) of other identities in a mini batch, and m is the margin to enhance discriminative ability of visual features. Note that, the each mini-batch contains N images randomly sampled from K identities. For faster convergence, an I_p^i (hardest positive) is selected such that $\text{argmax}_{I_p^i} \| \phi(I_a^i) - \phi(I_p^i) \|_2^2$ $, p = 1, \ldots, N'$ and similarly I_n^i (hard negative) such that $\text{argmin}_{I_n^i} \| \phi(I_a^i) - \phi(I_n^j) \|_2^2, n = 1, \ldots, N', j = 1, \ldots, K', j \neq i$. Intuitively, batch hard triplets helps the model to learn that similar looking images may belong to different identity (hard negatives) and images of the same person may appear differently (hard positives).

B. Local Feature Learning using Guided Second-order Attention Network

In person re-identification task, a set of unique or salient features of a person are crucial and identifying such salient features from a image of persons having similar geometry and appearance is a challenging task. For instance, to differentiate two persons, wearing similar clothes, it is important to consider their discriminative features like color of the clothes, type (e.g. full-sleeves, half-sleeves, shorts),

additional accessories (e.g. hand bags, carry bags), body pose and so on, which is difficult to notice even for human at first sight.

The majority of research efforts on person re-identification involves visual attention to extract salient features of an person and enhance feature representation[26], [27], [5], [28], [17], [6] . For example, [26], [27], directly locate the salient sub-regions using an attention network. [5], [17] proposes to integrate attention mechanisms to handle part alignment issue.[6] extend attention to channel dimension for refining feature representations. In addition, second-order attention methods [28], [29] are proposed to learn the higher order relationships in image spatial regions. Despite the developments, visual attention networks in person ReID cases still suffer several limitations. First, the visual attention mechanism is usually trained in weakly-supervised manner (i.e. no explicit labelling information is provided to identify the region to attend), which mean the attention maps learned this way are hardly corrected lacking discriminative ability and robustness. Second, these commonly used attention methods tend to capture only simple and coarse information due to their inherent first-order nature, ignoring higher-order statistical relationships of visual parts and the subtle differences among pedestrians. More importantly, majority of the works tend to apply attention at higher-level feature maps that are abstract. As a result, discriminating features of the person that can differentiate from other similar looking persons are typically ignored by such visual attention modules.

To overcome the aforementioned challenges and drawbacks, this paper propose a novel attention mechanism with two major improvement. First, we propose a second order self-attention module to capture fine-grained salient features and their second-order relationships at the early stages of CNN. Second, we propose a novel approach to guide our second-order self-attention module with a supervisory signal without any additional labelling information. More specifically, we train a evaluation network with attention maps of an image and use the output from the classifier to generate a loss value. This value is transmitted to the attention module using back-propagation for updating it's internal states, such that it improves the quality of the attention maps. With this self-attention learning approach, the attention module can efficiently determine where to attend, correct and adapt itself.

Let $\mathbf{x} \in \mathbb{R}^{w \times h \times c}$ denote the feature map of an arbitrary image I, where $h \times w$ is the spatial dimension and c is the number of channels of the feature map. With efficiency in mind, the channel dimension of $\mathbf{x}$ is reduced by a factor of r with a 1×1 convolution filter. Further, the $2-dimensional$ spatial maps is converted into single dimension yielding a tensor x with size $hw \times c/r$. The dimensionality reduction factor r is an important hyperparameter that reflect the amount of information pooled from the channel dimension to compute the attention weights. In our experiment this value is set to 2. The localization weights a for the feature maps x is computed as

$$a_i = \frac{\mathbf{x}_i W_1 W_2^T \mathbf{x}_i^T}{\sum_n^N \mathbf{x}_n W_1 W_2^T \mathbf{x}_n^T}, i = 1, \ldots, N \tag{4}$$

where $W_j \in \mathbb{R}^{c/r \times d}, j = 1, 2$ are learnable projection matrices. If we re-write $\mathbf{x}W_1W_2^T\mathbf{x}^T$ as $\mathbf{x}\Sigma\mathbf{x}^T$ where $\Sigma = W_1W_2^T$ then the numerator of equation (4) represents covariance matrix of x.Thus the attention weights $a \in \mathbb{R}^{wh \times wh}$ intuitively capture the correlations between each spatial positions of the feature tensor $\mathbf{x}$. The attention weights is once again linearly projected to c/r dimension resulting in final attention maps as

$$A(\mathbf{x}) = a \odot W_3\mathbf{x} \tag{5}$$

where $\odot$ is hardmard product between two tensors and W_3 is learnable matrix. Finally, we use a simple 1×1 convolution transformation to restore the channel dimension of the attention map A from c/r to c and define the second-order self-attention module as:

$$\mathbf{x} = \mathbf{x} + A(\mathbf{x}) \tag{6}$$

The attended feature maps $\mathbf{x}$ at layer L is then fed into the subsequent layer $L + 1$. In order to guide the attention learning, we design a 4-layer evaluation network in 'squeeze and excitation' manner. It is composed of a global average pooling layer and two linear projection layer. First, a *squeeze* operation is performed on the attention maps A via global average pooling operation to aggregate features across spatial dimensions.

$$A_s = \frac{1}{hw} \sum_i^h \sum_j^w A_{i,j,:c} \tag{7}$$

The *excitation* operation is formulated as

$$A_e = \sigma(W_2^e \times \max(0, W_1^e A_s)) \tag{8}$$

where $W_1^e \in \mathbb{R}^{c \times c/r}$, $W_2^e \in \mathbb{R}^{c/r \times c}$ are learnable matrix of the evaluation network, k is the number of classes. Note that excitation network with multi-step design doesn't increase model complexity, rather it reduces the number of parameters from c^2 to c^2/r by making the network narrower in width but longer in depth.

For attention maps belonging to the same person, we want the corresponding regions across different images of the person to be highlighted irrespective of the camera view, body pose and background. To achieve this, we enforce the attention network to be consistent in selecting salient features of a person from different feature maps using an attention loss as:

$$L_{att} = \alpha \sum_k^K \sum_n^N \parallel A_{e,a}^k - A_{e,p}^n \parallel_2^2 \tag{9}$$

The attention loss function aims to reduce the distance between the attention maps of the same person to the smallest possible value. α is the weighting factor for attention loss set to 0.1, $A_{e,a}^k$ is the attention maps of person a and $A_{e,p}^n$

is the positive pair of the same person. In order to determine the positive pair, we utilize the predicted probability of the true class labels from identity classifier. Since, each mini-batch contain N attention maps of a person a, we select one attention maps as a positive pair that has the maximum predicted probability of the true class label for the person a i.e $A^n_{e,p} := \text{argmax}_n\, p(y^n_a), n = 1, \ldots, N$, where $p(y^n_a)$ indicates the predicted probability of true class label for image n. By constraining the attention maps of a person to become as close as possible, the attention now focuses on the features that are common in all image view and different body pose of the person.

C. Tracklets Initialization and Association

The trajectory of a target k can be represented by a set of detections denoted by $T^k = \{b^k_1, b^k_2, \ldots\}$. The bounding box b^k_t of target k is defined by its coordinates (x_t, y_t, w_t, h_t), where x_t and y_t is the center location, w_t and h_t being the width and height of the bounding box respectively at frame t. Denote $D_t = \{b^1_t, b^2_t, \ldots, b^N_t\}$ as set of N detections at frame t which is provided by an off-line trained object detector. Multi-object tracking aim to estimate $T = \{T^k\}^K_{k=1}$ for all targets by assigning bounding box b^n_t to T_k.

At $t = 0$, our tracker initializes the tracklets from the set of N detections $D_0 = \{b^1_0, b^2_0, \ldots, b^N_0\}$ from the first frame. At each subsequent frames, the association between a detection $b^n_t \in D_t$ and a target $T^k \in T$ is measured to determine how well the detection is matched with the target as:

$$S(b^n_t, T^k) = S_a(b^n_t, T^K) S_m(b^n_t, T^K) \quad \forall n \in N, \forall k \in K$$

where S_a and S_m is the appearance and motion affinity between detections and targets respectively. The appearance affinity is measured as an euclidean distance between the learned CNN features from our appearance model while the Intersection of Union(IoU) between the bounding box detections and tracked targets is employed as the motion affinity. Note that, in a sequence, each D_t or T_k can have less elements than the actual trajectories or frames respectively. At each frame, a pairwise cost matrix $\mathbf{C} \in \mathbb{R}^{N \times K}$ is computed using Hungarian algorithm [30] from S for re-identification. Detection b^n_t for which the $S_{n,k} > \sigma_{sim}$ is added to the set of active tracks. Tracks that do not have any matching detections for time longer than $t_{inactive}$ is considered as inactive tracks. In order to account for new targets or reappeared targets, a new trajectory for the detection is initialized only if Intersection of Union (IoU) with any of the active trajectory is less than some threshold σ_{iou}. This ensure that the tracker will only consider the detections of potentially new object and suppress the overlapping detections at the same time.

D. Localizing the Lost Targets with Long-term Motion Prediction

Identity switching, assigning targets with incorrect labels, often occurs when the target detections are lost and reappeared after some interval. The problem is severe when the lost targets reappear in different location with new camera view. In our previous work [31], we discovered that motion prediction in long term can reduce identity switching problem and improves target recovery performance. When the targets are lost, we predict their trajectory for next t_{pred} frames using the prediction model [31] given their length of trajectory is greater than the observation time t_{obsv}. At each subsequent frames, any new detections that are not associated any active targets is matched with the predicted bounding boxes of the lost targets using IoU. The targets is considered recovered if the IOU between predicted bounding boxes and unmatched detection is greater than threshold σ_{iou}.

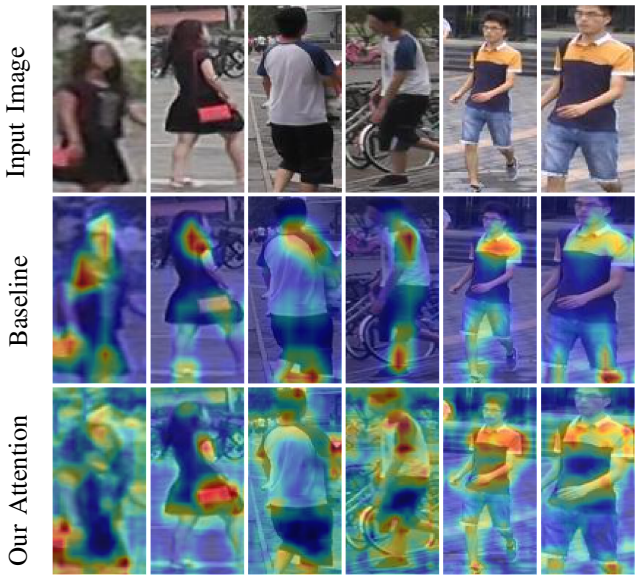

Fig. 3: Visualization of the attention maps from the baseline model and our self-guided attention learning. Each pair of columns, from left to right, corresponds to a single person. For each image, from top to bottom, we show the input image, baseline feature maps and our attended feature maps. As we can see, the baseline focus on few parts around body and legs, while our method give more attention to fine-grained features like bags, hair, hands, clothes and legs.

E. Data Augmentation

The performance of a deep convolutional network is affected by the number of training parameter and size of training samples. As the network become larger, it is easy to encounter over-fitting problem where the model prediction is very well on training data as compared to testing data. Explicit data augmentation technique such as random flipping, horizontal shift and random cropping helps to overcome over-fitting and improve the generalization ability of the model in unseen data. During training, we apply horizontal flip, random zoom, and random erasing [32] to the training samples. In random erasing, a random size mask (< 40 percent of image area) filled with uniform values is applied to fifty-percent of training images. Although random erasing [32] stimulate partial occlusion, which often occur in person re-identification, we note that this is not sufficient condition

TABLE I: Comparison of our ReID model performance with state-of-the-art methods on Market-1501 dataset.

Methods	Backbone	R-1	R-5	mAP
SVDNet [33] (2018 ICCV)	ResNet-50	82.3	92.3	62.1
CamStyle [16] (2018 CVPR)	ResNet-50	88.1	-	68.7
FD-GAN [34] (2018 NIPS)	ResNet-50	90.5	-	77.7
PCB [17] (2018 ECCV)	ResNet-50	93.8	97.5	81.6
SPReID [35] (2018 CVPR)	ResNet-152	93.7	97.6	83.4
SGGNN[36] (2018 ECCV)	ResNet-50	92.3	96.1	82.8
CAN [26] (2017 BMVC)	VGG-16	60.3	-	35.9
IDEAL [27] (2017 BMVC)	Inception-V3	86.7	-	67.5
HA-CNN [6](2018 CVPR)	HA-CNN	91.2	-	75.7
AACN [5] (2018 CVPR)	GoogleNet	88.69	-	82.96
MHN [7] (2019 ICCV)	ResNet-50	95.1	98.1	85.0
BAT [28] (2019 ICCV)	GoogleNet	95.1	**98.2**	**87.4**
Baseline	ResNet-50	93.08	95.72	81.94
GSAN	ResNet-50	**95.23**	97.46	85.42

to deal with occlusion and use it for the sake of fine tuning our model parameter.

F. Implementation Details

The appearance model was trained using SGD optimizer for 120 epochs with mini-batch consisting of randomly selected 16 identities, each with 4 images. The input image is resized to 256×128 and each color-channel is zero-centered with the mean values of ImageNet dataset. During evaluation, only resize and zero-centering of the input images is performed. The margin and weighting coefficient for the triplet loss is set to 0.3 and 0.0005 respectively. For tracking algorithm, the values of t_{obsv}, t_{pred}, $t_{inactive}$, σ_{iou}, σ_{sim} were set to 8, 12, 10, 0.2 and 0.5 respectively.

The learning rate is the one and most critical parameter to impact network convergence and subsequently model accuracy. A common approach to deal with network convergence is to use a constant learning rate at beginning and gradually decrease in regular intervals. An advantage of this method is higher convergence rate and reduced number of training epochs with comparable or even better results than that of with constant learning rate. Based on our experiment, we set a very small initial value η_{min} at the beginning of the training and gradually increase the value to η_{max} for 5 epochs and train for 30 epochs. For the remaining epochs, the learning rate is decayed by factor of 0.1 every 30 epochs. The learning rate η at each epoch is defined as:

$$\eta_T = \begin{cases} \eta_{min} + (\eta_{max} - \eta_{min})\frac{T}{5} & \text{if } T \leq 5 \\ \eta_{max} & \text{if } 5 < T \leq 30 \\ \eta_{max} \times 0.1^{\frac{T}{30}} & \text{if } T > 30 \end{cases} \quad (10)$$

IV. Experiments and Results

A. Datasets

We use MOT16 [37] and MOT17 [38] and Market-1501 [39] datasets for training and evaluation.

MOT16 includes trajectory sequences from popular datasets like KITTI, PETS09, TUD-Darmstadt and ETH-Crossing. It contain fourteen sequences (seven train, seven test)of different environment and various scenario like moving camera, densely crowded busy crossing and so on.

MOT17 contains total of 42 sequences (21 for training and 21 for testing). It is extensions of MOT16 with three detector: F-RCNN, SDP and and DPM detectors. The ground truth annotations of MOT16 and MOT17 are released whereas ground truth annotations of test sequences are not publicly available and used by the benchmark automatically for evaluation.

Market-1501 is considered as the largest person re-identification dataset containing 32,643 annotated bounding boxes of 1,501 identities collected from six cameras. The training set consists of 751 identities with 12,936 bounding images while the testing set consists of 750 identities in 19,732 images. During evaluation, a separate 3,368 images of 750 identities are searched in the test set to identify the correct identity. Our evaluation for this dataset is based on the single query method.

B. Evaluation Metrics

The performance of the tracking algorithm is evaluated with the widely adopted metric proposed by [38] including multiple object tracking precision (MOTP), multiple object tracking accuracy (MOTA) that combines Fasle Positives (FP), False Negatives(FN) and ID switches (IDS). We also report additional metric such as mostly tracked targets (MT), mostly lost targets (ML) and speed (Hz).

We use two evaluation metric to measure the performance of re-identification. The first metric is the Cumulated Matching Characteristics (CMC), which record the true matching within the top$-n$ results. Since the evaluation is positive even if there is one ground truth match in the gallery images for a given query, CMC is accurate only when one ground truth for each query exists and biased when multiple ground truth exists. To overcome the shortcomings, the second metric mean average precision (mAP) [39] is used. For each query, an Precision-Recall curve is calculated first which is averaged for all queries to compute mAP. As it consider both precision and recall, mAP provide comprehensive evaluation of the re-identification algorithm.

C. Re-Identification Results and Comparison with the State-of-the-Art Methods

We evaluate our re-identification model on Market-1501 and adopt single-query evaluation mode. Table I shows the Rank-1, Rank-5 and mAP values of the proposed approach and existing state-of-the-art methods. The bottom group summarizes the results of attention based re-identifications models. We consider ResNet-50 trained with cross-entropy and triplet loss as the baseline. It is clear from the comparison table that the proposed method achieves a superior results in terms of $R-1$ accuracy. In case of mAP, the proposed approach have competitive performance except for BAT that uses GoogleNet[40] as backbone network. We observe that adding second-order self attention from early stages of ResNet-50 improves the performances by $+4.24\%$ on mAP and 2.30% on Rank-1 accuracy from the baseline

TABLE II: Comparison of our tracking results with other state-of-the-art tracking methods on the test sets of MOT benchmark.

	Methods	MOTA ↑	MT↑	ML↓	FP↓	FN↓	IDS↓
MOT16	MHT_DAM [13](2015 ICCV)	45.8	16.2	43.2	6,412	91,758	**590**
	AMIR[15](2017 ICCV)	47.2	14.0	41.6	**2,681**	92,856	774
	MOTDT[14](2018 ICME)	47.6	15.2	38.3	9,253	85,431	792
	NOTA [41] (2019 ISL)	49.8	17.9	37.7	7,248	83,614	614
	Tracktor++ [11](2019 ICCV)	54.4	19.0	36.9	3,280	79,149	682
	Proposed	**57.3**	**28.1**	**24.1**	12,906	**63,590**	1,394
MOT17	MHT_DAM17 [13](2015 ICCV)	50.7	20.8	36.9	22,875	252,889	2,314
	jCC [42] (2018 PAMI)	51.2	20.9	37.0	25,937	247,822	**1,802**
	JBNOT[43] (2019 CVPR)	52.6	19.7	35.8	31,572	232,659	3,050
	FAMNet [44](2019 ICCV)	48.7	19.1	33.4	14,138	253,616	3,072
	Tracktor++ [11] (2019 ICCV)	53.5	19.5	36.6	**12,201**	248,047	2,072
	Proposed	**57.91**	**23.32**	**29.90**	24,831	**206,952**	5,667

performances. This highlights importance of self-attention in visual appearance models.

In order to validate the effectiveness of our self-attention model, the attention maps from the last conv5 layer is visually examined. Some illustration of the attention maps is shown in Figure 3. We choose two images of the same person with different pose and camera angle. As expected, the attention helps to focus on the salient and fine-grained features of the person despite variation in pose, viewpoint and background. We can observe that, discriminative components like bags, body parts, clothes etc. are highlighted in the feature map. This demonstrates that the proposed method is able to focus on unique features of the targets. Furthermore, we also compare the attention maps with the baseline feature maps. As we can see, the baseline features mostly focus on three regions only; head, body and legs of the targets irrespective of their salient attributes. This further verify that the proposed attention learning method proves effective in identifying salient regions and guiding appearance model to learn target specific features.

D. Tracking Results and Comparison with the State-of-the-Art Methods

The evaluation of the proposed tracking method is performed on the test set of the MOT16 and MOT17 challenge benchmarks. The overall comparison with state-of-the-art methods is summarized in Table II. The tracking results are obtained by replacing the re-identification model of [11] by GSAN. As in[11], we didn't not use any new detections other than the publicly available detections from the dataset and used their object detector only to compute confidence scores and perform non-maximal suppression of redundant detections. Moreover, the local features from the re-identification model was directly used for tracking without training on MOT benchmark. This allows us to verify the generalizability of our appearance model.

The comparison table shows two major benefits of using the proposed GSAN. First, significant improvement in tracking accuracy by $+5.33\%$ and 8.24% is achieved on MOT16 and MOT17 respectively. The performance gain in MOT17 is slightly higher than MOT16 since it is benefited SDP and FRCNN sequences which is not available in MOT16. Second, the ability to persistently track difficult targets and recover lost targets is achieved as can be seen from improved MT and ML. However, the number of FP is significantly higher despite the performances gain. This can occur when the threshold σ_{th} considered for matching detections and correct identity is very low allowing less similar images to be associated. Nevertheless, lowering threshold will also decrease FN - detections having low confidences is associated to the correct targets.

TABLE III: Effect of GSAN on MOT16 tracking performance (train set).

Models	MOTA ↑	MT↑	ML↓	FP↓	FN↓	IDS↓
Baseline	60.97	183	93	4566	37177	1347
c3	**61.13**	**187**	**90**	4597	**37106**	**1212**
c4	61.06	183	94	4542	37182	1264
c5	61.04	183	94	4542	37207	1264
c3,*c5*	61.06	184	93	4564	37191	1295
c3,*c4*,*c5*	61.07	184	92	**4534**	37196	1319

Effect of the position of GSAN: Table III shows the effect of GSAN when added to different stages of ResNet50. *c3, c4* and *c5* indicate the position of GSAN in ResNet50's stage 3, 4 and 5 respectively. The attention network can be inserted in multiple layers at the same time. Table III shows that: (1) using GSAN at early stages i.e. *c3*, *c4* proves to be more beneficial than using at later stages i.e. *c5*. (2) The performance of adding GSAN after *c3* alone surpasses the performance when it is added after *c4* and/or *c5*. One possible explanations is that the features maps at the earlier stages have richer information and spatial irregularities that can be easily captured by GSAN. Moreover, GSAN, regardless of the applied position, always improves tracking performance, however it has significant impact when placed at earlier stages. These findings are consistent with the observation in [28] where the bilinear attention enhances the person retrieval performance when placed at the early stages of GoogLeNet.

V. Conclusions

Multi-object tracking in crowded often face complex challenges like similar appearances of different identity, camera-view changes, occlusions and background clutter. Given the fact that similar looking person are not necessarily identical, we propose a second-order attention network where discriminative features from of a person is captured from context-rich features at earlier stages. Moreover, the attention learning methods are usually semi-supervised, we propose a novel Guided Second-order Attention Network (GSAN) which aim to achieve consistent feature learning despite of pose and view variation. Experimental results and analysis shows that proposed method achieve superior result on the challenging MOT17 benchmark.

Acknowledgement

This research is partially supported by SERC grant No. 162 25 00036 from the National Robotics Programme (NRP), Singapore.

REFERENCES

[1] Cao Qixin, Huang Yanwen, and Zhou Jingliang. An evolutionary artificial potential field algorithm for dynamic path planning of mobile robot. In *International Conference on Intelligent Robots and Systems*. IEEE, 2006.

[2] Kuffner Jr James J. and Steven M. LaValle. Rrt-connect: An efficient approach to single-query path planning. In *Robotics and Automation*. IEEE, 2000.

[3] David S. Bolme, J. Ross Beveridge, Bruce A. Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2010.

[4] Li Bo, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Computer Vision and Pattern Recognition*. IEEE, 2018.

[5] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[6] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE conference on computer vision and pattern recognition*, 2018.

[7] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *IEEE International Conference on Computer Vision*, 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on Computer Vision and Pattern Recognition*, 2016.

[9] Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Multi-person tracking by multicut and deep matching. In *European Conference on Computer Vision*. Springer, 2016.

[10] Bing Wang, Gang Wang, Kap Luk Chan, and Li Wang. Tracklet association by online target-specific metric learning and coherent dynamics estimation. *IEEE transactions on pattern analysis and machine intelligence*, 39(3), 2016.

[11] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *IEEE International Conference on Computer Vision*, 2019.

[12] Samuel S Blackman. Multiple hypothesis tracking for multiple target tracking. *IEEE Aerospace and Electronic Systems Magazine*, 19(1):5–18, 2004.

[13] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M Rehg. Multiple hypothesis tracking revisited. In *Computer Vision*. IEEE, 2015.

[14] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018.

[15] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *International Conference on Computer Vision*, 2017.

[16] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[17] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *European Conference on Computer Vision*, 2018.

[18] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017.

[19] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *IEEE conference on computer vision and pattern recognition*, 2018.

[20] Peng Li, Jiabin Zhang, Zheng Zhu, Yanwei Li, Lu Jiang, and Guan Huang. State-aware re-identification feature for multi-target multi-camera tracking. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[21] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[22] Anton Milan, Stefan Roth, and Konrad Schindler. Continuous energy minimization for multitarget tracking. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):58–72, 2013.

[23] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Computer Vision*. IEEE, 2009.

[24] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Computer Vision and Pattern Recognition*, 2016.

[25] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. *arXiv:1806.01482*, 2018.

[26] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7), 2017.

[27] Xu Lan, Hanxiao Wang, Shaogang Gong, and Xiatian Zhu. Deep reinforcement learning attention selection for person re-identification. *British Machine Vision Conference*, 2017.

[28] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *IEEE International Conference on Computer Vision*, 2019.

[29] Bryan Ning Xia, Yuan Gong, Yizhe Zhang, and Christian Poellabauer. Second-order non-local attention networks for person re-identification. In *IEEE International Conference on Computer Vision*, 2019.

[30] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics*, 52(1), 2005.

[31] N. Bhujel, E. K. Teoh, and W. Yau. Pedestrian trajectory prediction using rnn encoder-decoder with spatio-temporal attentions. In *International Conference on Mechatronics System and Robots*. IEEE, 2019.

[32] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

[33] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *IEEE International Conference on Computer Vision*, 2017.

[34] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in neural information processing systems*, 2018.

[35] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[36] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *European conference on computer vision*, 2018.

[37] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015.

[38] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[39] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE international conference on computer vision*, 2015.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition*, 2015.

[41] Long Chen, Haizhou Ai, Rui Chen, and Zijie Zhuang. Aggregate tracklet appearance features for multi-object tracking. *Signal Processing Letters*, 26(11), 2019.

[42] Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox, and Bernt Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):140–153, 2018.

[43] Roberto Henschel, Yunzhe Zou, and Bodo Rosenhahn. Multiple people tracking using body and joint detections. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[44] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *International Conference on Computer Vision*. IEEE, 2019.