Analysis of Recent Re-Identification Architectures for Tracking-by-Detection Paradigm in Multi-Object Tracking

Haruya Ishikawa[1], Masaki Hayashi[1], Trong Huy Phan[2], Kazuma Yamamoto[2], Makoto Masuda[2] and Yoshimitsu Aoki[1]

[1]*Department of Electrical Engineering, Keio University, Yokohama, Japan*

[2]*OKI Electric Industry Co., Ltd., Saitama, Japan*

Keywords: Multi-Object Tracking, Person Re-Identification, Video Re-Identification, Metric Learning.

Abstract: Person re-identification is a vital module of the tracking-by-detection framework for online multi-object tracking. Despite recent advances in multi-object tracking and person re-identification, inadequate attention was given to integrating these technologies to provide a robust multi-object tracker. In this work, we combine modern state-of-the-art re-identification models and modeling techniques on the basic tracking-by-detection framework and benchmark them on heavily occluded scenes to understand their effect. We hypothesize that temporal modeling for re-identification is crucial for training robust re-identification models for they are conditioned on sequences containing occlusions. Along with traditional image-based re-identification methods, we analyze temporal modeling methods used in video-based re-identification tasks. We also train re-identification models with different embedding methods, including triplet loss, and analyze their effect. We benchmark the re-identification models on the challenging MOT20 dataset containing crowded scenes with various occlusions. We provide a thorough assessment and investigation of the usage of modern re-identification modeling methods and prove that these methods are, in fact, effective for multi-object tracking. Compared to baseline methods, results show that these models can provide robust re-identification proved by improvements in the number of identity switching, MOTA, IDF1, and other metrics.

1 INTRODUCTION

Multi-Object Tracking (MOT) is one of the fundamental problems in computer vision that remain unsolved. Due to recent progress in object detection, tracking-by-detection emerged as the go-to framework in online MOT for its simple architecture and easy deployment. Generally, the tracking-by-detection framework consists of two parts: i.e., object detection and data association. We first detect objects, often people or automobiles, using an object detector that outputs bounding boxes and their confidence scores. A non-maximum suppression is employed to prune these detections and reduce false positives concentrated in a single area. We utilize a data association to connect the new detections to the history of identified detections, usually termed as tracklets. The simplest form of data association is to use IoU, proposed in the Simple Online and Realtime Tracking (SORT) (Bewley et al., 2016). SORT uses the history of detections in a tracklet to estimate the track's location in the next frame using Kalman Filter (Welch et al., 1995) and calculate the IoU of the estimations with the actual detections of the current frame. The IoU distances between the new detections and all estimated detections from existing targets from an assignment cost matrix. This assignment cost matrix is then solved using the Hungarian algorithm (Kuhn, 1955). SORT has become a baseline method for simplicity and runtime. However, SORT suffers from tracking lost and identity switching (IDS) due to a heavy dependency on the detector's accuracy and occlusions. Tracking loss and IDS have been a significant challenge for MOT, and many complex methods tried to solve this issue. Deep SORT (Wojke et al., 2017) builds on top of SORT to solve these issues by incorporating a re-identification (Re-ID) model as a feature extractor to extract feature vectors of the detections. These feature vectors are used to match already confirmed targets against new detections to re-identify occluded targets that are temporally lost. In this matching cascade, the cosine distances of the feature vectors are used in an assignment cost matrix. With appearance information, Deep SORT has proven

Ishikawa, H., Hayashi, M., Phan, T., Yamamoto, K., Masuda, M. and Aoki, Y.
Analysis of Recent Re-Identification Architectures for Tracking-by-Detection Paradigm in Multi-Object Tracking.
DOI: 10.5220/0010341502340244
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 234-244
ISBN: 978-989-758-488-6

to re-identify occluded objects by lowering IDS and loss of tracking.

Recently, great progress has been made in the field of Person Re-ID. Person Re-ID is a prevalent computer vision task consisting of re-identifying, or querying, images of a person in an extensive gallery of people. Many advances have been made in person Re-ID, which include various models for feature extraction and training methods for robust embedding network. As people are naturally the center of most computer vision tasks, it also holds the same for MOT. For the goal of tracking people, MOTChallenge hosts numerous MOT competitions and datasets (Leal-Taixé et al., 2015; Milan et al., 2016; Dendorfer et al., 2020b). Deep SORT uses a simple person Re-ID model architecture based on ResNet and is trained on a large Re-ID dataset. Many recent MOT trackers that use a Re-ID model use a similar simple architecture based on ResNet. However, simple Re-ID models become inadequately discriminative as MOTChallenge began using very crowded scenes with low detection confidences and various people-to-people or people-to-object occlusions like the MOT20 dataset (Dendorfer et al., 2020b). This study explores the possibility of better performance in MOT by modern Re-ID models. We investigate modern Re-ID methods on the task of MOT and only use MOT dataset to trian the Re-ID models.

We believe that the full potential of Re-ID models for MOT are yet to be discovered. In this paper, we extensively study the effects of applying state-of-the-art person Re-ID methods to the task of MOT. More specifically, we train these Re-ID models for the data association module of the tracking-by-detection paradigm, with the baseline being Deep SORT. The contributions made in this paper are as follows:

- We exploit state-of-the-art person Re-ID models for our tracker and compare them against baseline DeepSORT for the challenging MOT20 dataset. The only dataset used for training these Re-ID models is the MOT20 dataset.
- We analyze the effects of metric learning methods in order to create more robust embedding feature vectors. Our work investigates the effects of verification loss and triplet loss along with simple identity loss.
- We apply video person Re-ID training methods such as temporal attention and evaluate the effects they have on the embedding feature vectors for Re-ID in the MOT task.

We hope these comparisons of Re-ID models serve as important insights for new Re-ID architectures for tracking-by-detection framework in the future.

2 RELATED WORKS

Multi-Object Tracking (MOT). Multi-object tracking (MOT) aims at predicting trajectories of multiple targets in video sequences. It is crucial for various computer vision applications such as surveillance, activity recognition, and autonomous driving. One of the most popular targets is people because of the diversity in their looks and behaviors. MOTChallenge (Leal-Taixé et al., 2015; Milan et al., 2016; Dendorfer et al., 2020b; Dendorfer et al., 2020a) is an open competition where institutes compare their trackers on various MOT metrics on video sequences depicting movements of people. Despite the many works in MOT, re-identifying and reducing identity switching remains a challenging problem, especially in crowded scenes where occlusions and false alarms are common.

Multi-object tracking is commonly classified as either online or batch methods. Online methods can only use current and previous tracked frames, whereas batch methods can use the entire sequence. The class of batch methods typically perform better than online methods since they can utilize the whole sequence and solve a global optimization problem. However, more applications such as surveillance and autonomous driving need realtime performance. Thus, we assume online methods in this paper. Tracking-by-detection is a dominant strategy for online MOT, which treats detection and data association as two separate tasks (Bewley et al., 2016; Yu et al., 2016; Wojke et al., 2017; Choi, 2015). Recent deep learning based MOT methods apply CNN-based object detectors such as Faster R-CNN (Ren et al., 2015) and YOLOv3 (Farhadi and Redmon, 2018) to localize all objects of interest in the input image. Data association is employed in the next step, where the detections are linked with previous detections to create tracklets. Most of these methods utilize an identity embedding network for re-identification in the data association step, commonly known as the Re-ID model. Following SORT and Deep SORT's tracking-by-detection framework, it is standard practice that, we first compute an assignment cost matrix according to the Re-ID features and IoU of the bounding boxes and then use the Kalman filter (Welch et al., 1995) and Hungarian algorithm (Kuhn, 1955) to accomplish the association. Some variants are also proposed, which use more complicated association strategies such as RNNs and group models (Fang et al., 2018; Zhou et al., 2018; Mahmoudi et al., 2019). The main advantage of using tracking-by-detection methods is that they can develop the most suitable model for each task separately without making compromises. Moreover,

they can also crop the image patches according to the detected bounding boxes and resize them to the same size before estimating Re-ID features. Cropping the images helps to handle the scale variations of objects. As a result, these approaches have achieved the best performance on public datasets, nevertheless heavily relying on the underlying detection method's performance. In this work, we focus on the data association, specifically the Re-ID model, without going into deep discussions of the detection problem. Therefore, we use a public detector given by the MOT datasets instead of the private detectors used in various state-of-the-art methods to benchmark the Re-ID models fairly.

Recent methods on one-shot tracker has been very successful (Wang et al., 2019; Zhan et al., 2020). These methods treat MOT as a multi-task learning problem that try to simultaneously accomplish object detection and identity embedding in a single network in order to reduce inference time. One-shot trackers such as JDE (Wang et al., 2019) and FairMOT (Zhan et al., 2020) jointly train the detector and identity embedding using the same backbone and a variety of heads for trainining the backbone. We do not address one-shot trackers directly in our work since they train their own private detectors, yet we believe that the backbone used in these methods can adopt similar backbones used in the Re-ID method. We believe our analysis of modern Re-ID models in MOT could bring useful insights into choosing these backbones.

Person Re-Identification (Re-ID). Person Re-ID, or image person Re-ID, is a common computer vision task consisting of re-identifying, or querying, images of a target person from a gallery. In person Re-ID, a model must extract feature vectors that are discriminative enough to differentiate among different people but similar enough for the model to identify the targets. Therefore, a robust Re-ID model should have a robust embedding vector space. With the technical advancements of CNNs, the backbone of CNNs is used to facilitate multi-dimensional embedding. There exist various person Re-ID datasets to train these Re-ID, such as Market1501 (Zheng et al., 2015). It is very common for these datasets to have cleanly prepossessed detections, which are very similar to MOT's data association problem. Our work investigates the effects of state-of-the-art person Re-ID models such as OSNet (Zhou et al., 2019) and RGA-SC (Zhang et al., 2020) for Re-ID in MOT.

Similarly, video (person) Re-ID is also a common computer vision task, which uses a sequence of images instead of a single image for re-identifying. Compared to image person Re-ID, video sequences often portray the target's occlusions, adding challenges to re-identification. Various occlusions occur in MOTChallenge and more so in the recent MOT20 dataset. Therefore, video Re-ID models are useful for MOT for they are conditioned on these scenes and can extract robust features.In our work, we train the Re-ID models that are conditioned on sequential data much like video Re-ID models. We also analyze the effects of temporal training methods such as RNN (McLaughlin et al., 2016) and temporal attention (Liu et al., 2017) on Re-ID models for MOT.

3 PROPOSED METHOD

In this work, we build on top of the tracking-by-detection framework to deal with crowded scenes with heavy occlusion like the sequences in MOT20. To overcome this problem, we utilize the state-of-the-art person Re-ID models and various video Re-ID training methods. As most of the current Re-ID models used in MOT are variants of ResNet (He et al., 2016), we use ResNet50 as our baseline Re-ID model. In Section 3.1, we briefly introduce the state-of-the-art Re-ID models that are considered for our analysis. In Section 3.2, we propose video Re-ID methods and models for conditioning the Re-ID backbone on occluded scenes. In person Re-ID, various metric learning methods are being studied to embed the features more robustly. In Section 3.3, we go over the simple identity loss along with verification loss and triplet loss. Finally in Section 3.4, we examine Deep SORT (Wojke et al., 2017), the tracker that we use throughout our experiments.

3.1 Image Re-ID Models

With the popularity of person Re-ID, current Re-ID models are considerably improved. We have selected the state-of-the-art Re-ID models that have achieved competitive results in various datasets; more specifically PCB+RPP (Sun et al., 2018), OSNet (Zhou et al., 2019), RGA-SC (Zhang et al., 2020), and CBAM (Woo et al., 2018). In this section, we will explain the mechanics of these models.

PCB and RPP. Part-based Convolutional Baseline (PCB) was introduced in (Sun et al., 2018). The model architecture consists of the same ResNet50 for the backbone. However, instead of using the Global Average Pooling (GAP) for pooling, PCB pools from n divided regions into n features, as shown in Figure 1. Compared to the baseline (ResNet50) model, PCB benefits from having n distinct part-level features because these features offer fine-grained information and could lead to learning part-informed fea-

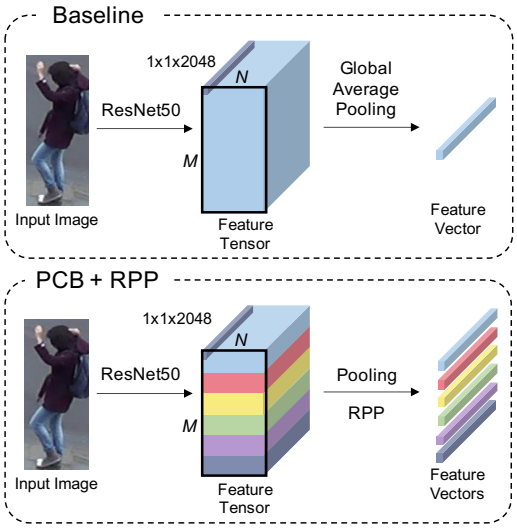

Figure 1: Baseline model and PCB+RPP.

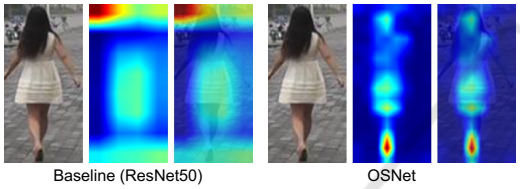

Figure 2: Heatmap representation of feature tensor for baseline model and OSNet on Mars dataset. OSNet can capture discriminative features such as the shoe the woman is wearing, the pattern on the skirt, and the global shape of the woman, whereas baseline model suffers from an over focus on the background on the top left without capturing robust features.

tures. Refined Part Pooling (RPP) is also proposed to reinforce the within-part consistency in each part instead of taking an average pooling approach. After the refinement, similar column vectors are concluded into the same part, making each part more internally consistent. We utilize PCB along with RPP, naming the model PCB+RPP throughout the paper. n is set to 6 for all our experiments.

OSNet. Omni-scale network (OSNet) is coined in (Zhou et al., 2019), which completely re-designs CNN to capture both homogeneous and heterogeneous scaled features, with homogeneous scale being in the sense of global features such as body shape, clothing, etc, and heterogeneous scale meaning combinations of richer features spanning in multiple scales such as combination of clothing, body attributes, etc. OSNet can capture relatively discriminative attributes compared to the baseline model, as shown in Figure 2. It is noted here that we use the best performing width multiplier $\beta = 1.0$ and resolution multiplier $\gamma = 1.0$ for all our experiments.

Table 1: Results on Market1501 Dataset. The only loss used is the ID loss (I).

Model	Loss	mAP	CMC1
Baseline	I	69.94%	86.76%
PCB+RPP		72.84%	87.84%
OSNet		65.6%	78.21%
CBAM		70.07%	85.86%
RGA-SC		72.44%	87.91%

Attention-based Methods. With the recent success in applying attention mechanisms to CNN, we consider CBAM ResNet50 and RGA-SC ResNet50, proposed in (Woo et al., 2018) and (Zhang et al., 2020) respectively. They are both modifications of the baseline model and add channel- and spatial-wise attention blocks between the ResNet bottleneck blocks. CBAM block focuses more on the local attention, whereas RGA-SC block focuses on the global attention of the features. For convenience, we term CBAM ResNet50 as CBAM and RGA-SC ResNet50 as RGA-SC throughout the experiments.

Ablation Studies on Market1501. Although official and unofficial implementations of these models are available online, we re-implement them according to their implementations to reproduce their results. We modify the architectures after GAP (or RPP) and added a single MLP layer for the classification head for fair comparisons. We train and test the models on the Market1501 dataset (Zheng et al., 2015). For training, we use pretrained weights trained on ImageNet (Deng et al., 2009) for ResNet variants, as well as Kaiming initialization where pretrained layers were not available. A simple ID loss with label smoothing is used. Adam optimizer is used with a weight decay of 0.0005 and a learning rate of 0.0003. We use a single step learning rate scheduler with the step being at 20 epoch and decreases learning rate by 10^{-1}. We trained 200 epochs in total for all the models and used the best metrics for the results which we show in Table 1. We understand that bag-of-tricks (Luo et al., 2019) for improving each model exist, but in this paper, we focus on a fair comparison of the models under the same conditions. Most of the models are performing correctly but not close to the metrics proposed in the papers under these conditions.

3.2 Video Re-ID Methods

Video Re-ID, as explained in Section 2, tackles Re-ID for sequential data that depicts subtle to major visual changes that include occlusion and crowded environments. In video Re-ID, a model must compensate for these occlusions or view changes and perform robust Re-ID. It can be said that video Re-ID models are conditioned on these kinds of data, thus more ro-

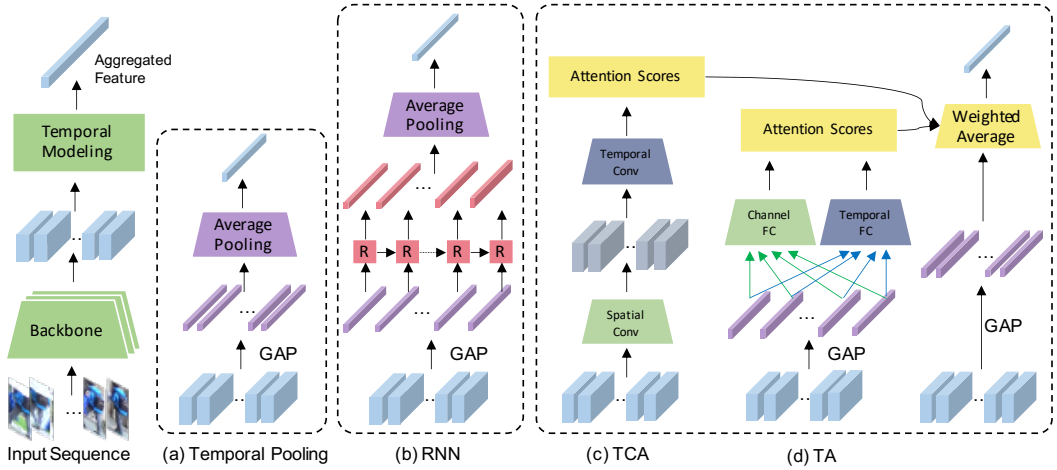

Figure 3: Visualization of temporal models used in video Re-ID. (a) Most naive method is to use temporal (average) pooling. (b) RNNs are utilized to take advantage of the sequential nature of video Re-ID. (c) Temporal Convolutional Attention (TCA) uses spatial and temporal convolutions to retrieve attention scores for weighted averaging of features. (d) Temporal Attention (TA) creates attention scores through channel and temporal fully connected (FC) layers.

bust than modeling used in image Re-ID. In Figure 3, we describe common video Re-ID modeling methods. We believe that it is crucial to investigate video Re-ID methods since most detections in MOT20 contain heavy occlusions. It is not feasible to train models using the image Re-ID method since most images would have to be pruned out due to occlusion, making the dataset very small. Training with video Re-ID methods can utilize the entire MOT training dataset without significant drawbacks.

Common Temporal Modeling. A naive temporal modeling method is to directly use average pooling of all the features gathered separately through the backbone, as is depicted as temporal pooling (TP) in Figure 3. Another method is using RNN to model the sequential nature of the video. In our work, instead of taking the last RNN feature, we use the average pooling of all the intermediate features.

Attention-based Temporal Modeling. Attention-based temporal modeling is commonly used in video Re-ID to mask the features that can potentially gather robust features for Re-ID. Essentially, if the sequence contains heavy occlusions, those frames would have lower attention scores, making use of the more confident features. Our work investigates temporal convolutional attention (TCA), which also takes advantage of the spatial convolution for attention modeling, and temporal attention (TA).

Ablation Studies on Mars. Following the works of (Gao and Nevatia, 2018), we re-implement temporal modeling methods. We evaluate temporal methods along with image Re-ID models on the Mars dataset (Zheng et al., 2016), as shown in Table 2. Each model is trained in the same manner as the results in Table 1, except for that each model uses a gradual learning rate warm-up until 20 epoch. It uses a multi-step learning rate scheduler with a step size of 50 epochs that decreases the learning rate by 10^{-1}. We train the RNN with the initial learning rate of 0.00015. From the table, we can see that OSNet performs relatively well, considering that it does not perform well in Table 1. The cumulative matching characteristics (CMC) rank-1 metrics for temporal modeling methods are notably high.

Table 2: Results on Mars Dataset. Losses used are ID loss (I), verification loss (V), and triplet loss (T). Temporal modeling techniques are temporal pooling (TP), recurrent neural network (RNN), temporal attention (TA), and temporal convolutional attention (TCA).

Model	Loss	Temp.	mAP	CMC1
Baseline	I	TP	62.65%	74.24%
PCB+RPP			61.98%	71.74%
OSNet			72.70%	77.54%
CBAM			73.16%	79.89%
RGA-SC			66.88%	75.25%
Baseline	I	RNN	65.65%	79.24%
		TA	69.93%	79.35%
		TCA	71.31%	80.43%
Baseline	I+V	TP	30.45%	40.27%
	I+T		71.65%	79.46%
	I+V+T		44.34%	55.6%

3.3 Metric Learning

ID Loss. Each person is labeled by ID and is treated as a classification problem when we are training with ID loss. We assume that the x_i and y_i are i-th frame and identity label. The probability of x_i being predicted as y_i can be encoded by a softmax function and

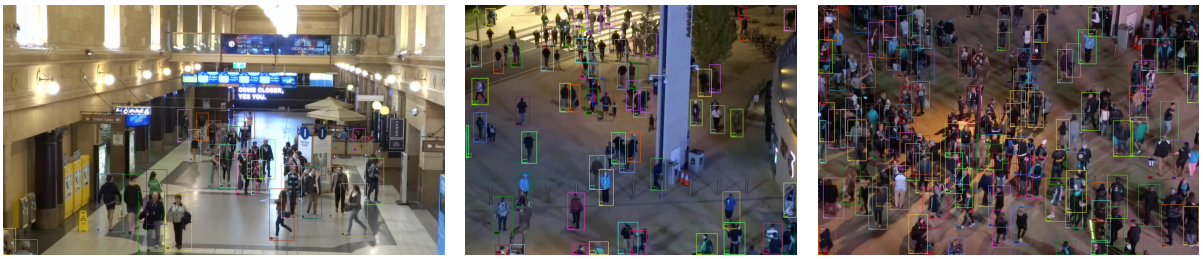

Figure 4: Visualization of MOT20 dataset's training set.

is represented as $p(y_i|x_i)$. ID loss is a cross-entropy loss and is expressed as,

$$L_{id}(i) = -\frac{1}{n}\sum_{i=1}^{n}\log(p(y_i|x_i)) \quad (1)$$

where n is the batch size. To ensure that the model doesn't overfit to the randomly annotated label, we employ label smoothing (Zheng et al., 2017b).

Verification Loss. Given a pair of images, we use binary cross entropy to classify whether it is the same person or a different person in a binary verification loss (Zheng et al., 2017a). This can be expressed as,

$$L_{ver}(i,j) = -\delta_{ij}\log(p(\delta_{ij}|f_{ij})) - (1-\delta_{ij})\log(1-p(\delta_{ij}|f_{ij})) \quad (2)$$

where f_{ij} is a differentiable value expressed as $f_{ij} = (f_i - f_J)^2$ and δ_{ij} is a binary label ($\delta_{ij} = 1$ when x_i and x_j is the same identity).

Triplet Loss. In a triplet loss, we form a triplet of anchor x_i, positive x_j (same identity as anchor), and negative x_k (different identity from anchor). We embed features closer for the same identities and, at the same time, embed features further apart when the identities are different. If ρ is the margin constant of the embedding space, we can express the triplet loss as,

$$L_{tri}(i,j,k) = \max(\rho + d_{ij} - d_{ik}, 0) \quad (3)$$

where $d(\cdot)$ represents the euclidean distance of the samples. We follow the work of (Hermans et al., 2017) and use batch hard triplet loss in order to mine the hardest triplet for the loss and minimize the effect of easier triplets. ρ is set to 0.3 in all our experiments.

Ablation Studies on Mars. In our work, we integrate all the losses into a single loss expressed as,

$$L = \frac{\alpha}{N_{id}}\sum_{i,j,k} L_{id} + \frac{\beta}{N_{ver}}\sum_{i\neq j, i\neq k, j\neq k} L_{ver} + \gamma L_{tri} \quad (4)$$

where (α,β,γ) are constants ($\alpha+\beta+\gamma=1$) and N_{id} and N_{ver} are the total sample size of each of the elements. On Table 2, we show results of using combinations of each of the metric learning losses. The value of (α,β,γ) for each of the combinations are: $(0.8, 0.2, 0)$ for I+V, $(0.8, 0, 0.2)$ for I+T, and $(0.6, 0.2, 0.2)$ for I+V+T. It is clear that adding verification loss does not improve, but rather degrade the metrics compared to only using ID loss. Adding triplet loss does improve the metrics of the baseline by 9% for mean average precision (mAP).

3.4 Tracking and Re-ID Model

We use Deep SORT as our base tracker throughout the experiments (Wojke et al., 2017). We choose Deep SORT for its simplicity, and it allows for a fair evaluation of the effects of each of the Re-ID model on the task of MOT. Continuing off of the explanation in Section 2, various parameters of Deep SORT must be discussed. From the detector, we obtain bounding boxes and their confidence scores. We put a threshold on the confidence scores to reduce the amount of false alarms. When the detections are unmatched to any of the prior tracks, new track hypotheses are created. The track hypothesis is in a tentative state until it is confirmed as a tracklet by having τ_{init} frames of matched detection. When hypotheses or tracklets have no new detections for τ_{max} frames, the tracks are considered to have left the scene and are in a deleted state. For data association, we keep a gallery of features for each track and save up to L_{max} features to match against. In our experiments, $\tau_{init} = 3$, $\tau_{max} = 70$, $L_{max} = 100$, and we use cosine distance as our distance metric (Wojke and Bewley, 2018). Since people are shown very far in the MOT20 dataset, most of the confidence scores are around 0. Therefore we reduce the threshold for confidence scores to 0.

4 EXPERIMENTS

MOT20 Dataset. For evaluating MOT, we use the MOT20 dataset provided by MOTChallenge (Dendorfer et al., 2020b). Since the number of submissions is limited, as advised by MOTChallenge, we use the training set of the dataset for our initial evaluation discussed in Section 5.1. More specifically, in MOT20,

Table 3: Upper-bound results of each video-based Re-ID method used in DeepSORT for sequence 05 of MOT20 Dataset. These upper-bound results are obtained from using the ground truth tracks as detections for each frame, which are associated using Kalman Filter and Re-ID module. These results provide a better understanding of how each Re-ID models perform. The "Vis Ratio" is the visible ratio of the training dataset used to train the Re-ID models. All the models are trained with ID loss and batch hard triplet loss.

Re-ID	Vis Ratio	MOTA ↑	IDF1 ↑	MOTP ↓	MT ↑	ML ↓	FP ↓	FN ↓	Rec. ↑	Prec. ↑	IDS ↓	Frag. ↓
BL	0.3	64.35%	60.55%	9.97%	645	105	2272	264505	64.78%	99.54%	1073	3001
	0.0	64.37%	59.42%	9.96%	646	104	2281	264331	64.82%	99.53%	1102	3020
CBAM	0.3	64.36%	59.12%	9.96%	644	106	2302	264392	64.81%	99.53%	1112	3000
	0.0	64.36%	59.68%	9.96%	646	105	2322	264417	64.81%	99.53%	1038	3002
OSNet	0.3	64.37%	56.84%	9.96%	645	103	2311	264102	64.85%	99.53%	1259	3017
	0.0	64.39%	57.14%	9.97%	649	104	2276	264042	64.86%	99.54%	1201	3021
PCB+RPP	0.3	64.44%	61.65%	9.95%	649	105	2209	264919	64.87%	99.55%	1032	2934
	0.0	64.08%	60.60%	9.92%	638	107	2596	266015	64.59%	99.47%	1279	2885
RGA-SC	0.3	63.91%	59.62%	9.93%	634	106	2640	267144	64.44%	99.46%	1365	2909
	0.0	64.02%	61.80%	9.91%	641	104	2605	266551	64.52%	99.47%	1176	2912

there are four training video sequences: 01, 02, 03, and 05. Since sequences 01 and 02 are taken from the exact point of view, and 03 and 05 differ greatly (as shown in Figure 4), we use 01, 02, and 03 for the training sequences and 05 as our test sequence. Note that sequence 05 has more population density, thus more occlusion than the other three sequences. Later, we use the best-performing methods to benchmark on MOTChallenge's server.

Training the Re-ID Models. As explained in Sections 3.1 and 3.2, two different kinds of Re-ID training methods are commonly used: image-based and video-based Re-ID. For the task of image-based Re-ID, the images are most likely to be clear and do not contain images with people that are heavily occluded. On the other hand, heavily occluded sequences often occur in the task of video-based Re-ID. Therefore, we refine the training set of the MOT20 dataset in training our models. More specifically, we take the ground truth detections of the MOT20 dataset and use them as our training dataset for our Re-ID models. For training image-based Re-ID models, we use detections with a visible ratio over of 0.3, which is a common practice in the previous MOTChallenges (Bergmann et al., 2019).

For training video-based Re-ID models, we use detections with a visible ratio of 0, which is the entire detection sequence of the MOT20 dataset. Since visible ratio for video-based Re-ID models has not been studied for the MOT20 dataset, we study an upper-bound evaluation on sequence 05 of the MOT20 dataset. Results on Table 3 show that a visible ratio of over 0 is acceptable for training our video-based Re-ID models. As for the training details, we randomly sample 8 consecutive frames from the tracklet to create our models' input.

Training details for all Re-ID methods for the MOT20 dataset are the same as the ablation studies for the Mars dataset results. We train each model for 200 epochs and use the last epoch for all our evaluations in Section 5 because we can not afford to create a validation set due to the limited training dataset available. Note that for benchmarking on MOT20 (Section 5.2), we use the entire training sequences to train our Re-ID models.

5 RESULTS

5.1 Evaluation on Sequence 05

Since we cannot evaluate all the combinations on the MOTChallenge benchmarking server, we will first evaluate our models on sequence 05 of the MOT20 dataset as shown in Table 4. We have evaluated both image and video Re-ID models. Image Re-ID models are trained on the image Re-ID dataset created from MOT20 public detections. Video Re-ID models are trained on video sequences of tracklets containing various occlusions. All Re-ID models are trained on sequences 01, 02, and 03. The only dataset used to train these models is the MOT20 dataset. Along with Deep SORT, we have included outputs of SORT, which is often the minimum standard in MOTChallenge. Results are colored according to the rankings (best seen on PDF).

Metrics used in the results are: MOTA, IDF1, MOTP, Identification Precision (IDP), Identification Recall (IDR), Recall (Rec.), Precision (Prec.), Mostly Tracked (MT), Partially Tracked (PT), Mostly Lost (ML), False Positives (FP), Misses (Miss), Identity Switching (IDS), and Fragmentation (Frag). Details of the definitions of each metrics are explained in (Dendorfer et al., 2020a).

Image vs Video Re-ID Dataset. Since the image training set created from the MOT20 dataset consists of redundant images of each identity from the high

Table 4: Results on sequence 05 of the MOT20 dataset. For Deep SORT, both "Image" and "Video" methods of training are used. 'Red" and "magenta" mean first and second-best values respectively and "blue" represents the worst value. As it is obvious from the colors, modern Re-ID methods are very advantageous over SORT. Models such as RGA-SC and PCB+RPP trained over video data using ID (I) and triplet (T) loss have many competitive edges over the Baseline (BL) model. One of them being the lower ID Switching (IDS) value, which has greatly reduced. It is also noted that besides OSNet, the backbone architecture for all models is ResNet50.

Tracker		Model	Loss	MOTA ↑	IDF1 ↑	MOTP ↓	IDP ↑	IDR ↑	Rec. ↑	Prec. ↑	MT ↑	PT ↑	ML ↓	FP ↓	Miss ↓	IDS ↓	Frag. ↓
SORT				45.18%	27.40%	13.16%	42.65%	20.19%	47.07%	99.44%	159	750	302	1991	397671	12209	14884
Deep SORT	Image	BL	I+T	51.68%	46.10%	13.49%	66.52%	35.28%	52.60%	99.19%	233	738	240	3238	356141	3632	17040
		CBAM	I+T	51.83%	43.39%	13.62%	62.36%	33.27%	52.88%	99.11%	236	747	228	3571	354042	4306	17664
		OSNet	I	51.95%	42.40%	13.59%	60.88%	32.53%	52.98%	99.15%	241	742	228	3396	353276	4308	17607
		PCB+RPP	I+T	52.00%	28.82%	14.00%	41.16%	22.17%	53.40%	99.14%	246	743	222	3487	350138	7038	18456
		RGA-SC	I+T	51.84%	27.96%	14.02%	39.95%	21.50%	53.32%	99.08%	247	744	220	3713	350703	7462	18446
	Video	BL	I	52.03%	38.76%	13.69%	55.56%	29.76%	53.10%	99.15%	246	742	223	3405	352376	4629	17871
		BL	I+T	52.01%	31.08%	13.86%	44.43%	23.90%	53.31%	99.13%	249	741	221	3528	350778	6285	18218
		BL	I+V	51.81%	27.89%	14.01%	39.83%	21.45%	53.36%	99.09%	247	744	220	3688	350413	7970	18630
		BL	I+V+T	51.83%	27.43%	14.00%	39.18%	21.10%	53.37%	99.10%	247	744	220	3650	350322	7973	18615
		BL+RNN	I	52.00%	32.99%	13.80%	47.22%	25.35%	53.23%	99.15%	245	746	220	3433	351411	5828	18095
		BL+RNN	I+T	51.97%	28.13%	13.97%	40.17%	21.64%	53.39%	99.11%	244	747	220	3590	350225	7064	18490
		BL+TCA	I	52.04%	39.23%	13.73%	56.23%	30.12%	53.11%	99.14%	236	751	224	3450	352314	4543	17908
		BL+TCA	I+T	52.06%	42.22%	13.64%	60.55%	32.41%	53.06%	99.14%	244	743	224	3457	352677	4077	17795
		BL+TA	I	52.03%	39.71%	13.67%	56.94%	30.49%	53.10%	99.15%	243	744	224	3415	352368	4634	17815
		BL+TA	I+T	52.06%	32.44%	13.79%	46.38%	24.95%	53.32%	99.12%	250	739	222	3575	350735	5883	18289
		CBAM	I	51.91%	33.56%	13.81%	48.05%	25.78%	53.18%	99.10%	242	746	223	3624	351762	5950	18130
		CBAM	I+T	52.15%	39.60%	13.65%	56.71%	30.41%	53.18%	99.16%	245	739	227	3389	351782	4377	17885
		OSNet	I	51.81%	29.66%	13.91%	42.41%	22.80%	53.26%	99.09%	246	743	222	3667	351158	7222	18400
		OSNet	I+T	52.11%	36.54%	13.69%	52.26%	28.09%	53.28%	99.12%	249	739	223	3571	351044	5183	18193
		PCB+RPP	I	51.72%	29.31%	13.89%	41.93%	22.53%	53.23%	99.05%	244	745	222	3817	351371	7588	18373
		PCB+RPP	I+T	51.07%	46.06%	13.36%	67.07%	35.08%	51.92%	99.27%	230	733	248	2879	361245	3526	16239
		RGA-SC	I	51.94%	37.26%	13.63%	53.43%	28.60%	53.06%	99.13%	238	751	222	3510	352655	4891	17817
		RGA-SC	I+T	51.43%	47.87%	13.37%	69.46%	36.52%	52.19%	99.27%	233	735	243	2902	359226	2785	16438
		CBAM+TA	I	51.94%	34.21%	13.79%	48.98%	26.28%	53.19%	99.14%	240	747	224	3483	351713	5911	18012
		CBAM+TA	I+T	52.09%	38.43%	13.66%	55.04%	29.52%	53.18%	99.15%	241	748	222	3438	351767	4753	17921

frame-rate, most of the modeling methods have failed to converge within the limited epoch. The results on the table are the ones that have managed to converge. Note that most models that have converged are ID and triplet loss. As only one tracklet exists for a single identity, the video training set is relatively small. Additionally, as a result of random sampling from the tracklets, it was much easier to form a batch, making the video Re-ID modeling approach easier to converge. All of the models have converged within the limited epoch. The best performing image Re-ID model is PCB+RPP, which performs 52.00% for MOTA, has the highest recall with 53.40%, and lowest misses of 350138. However, this comes with the cost of having one of the largest counts of IDS and fragmentation (i.e., the number of times the ground-truth trajectory is untracked). Video Re-ID, however, has the highest MOTA performance with CBAM (I+T), scoring 52.50% while maintaining lower IDS and fragmentation, which is commonly known to be a trade-off.

Effects of Metric Learning. Considering losses of baseline methods (in video Re-ID methods), we can see that the baseline method with ID loss and ID with triplet loss performs better than those with verification loss. This is of no surprise since verification loss performs poorly on the Mars dataset in Table 2. Models trained with verification loss have the lowest mostly lost values, yet at the costs of lower IDS, fragmentation, MOTA, and IDF1. It is hard to observe the effects of triplet loss by only looking at the baseline models. However, when we compare ID loss and triplet loss throughout the table, we can see significant improvements with various metrics such as fragmentation, IDS, IDF1, IDP, and IDR. Generally, we can observe that ID loss with triplet loss does not deteriorate the results and improve certain metrics.

Impact of Video Re-ID Modeling. Comparing the baseline model (baseline with ID loss) to baseline methods that use temporal modeling in video Re-ID models, we can see that TCA and TA have higher MOTA. TA (I+T) has the highest mostly tracked counts, low mostly lost counts, and a low number of misses. TCA (I) has the highest partially tracked counts with low mostly lost counts and IDS. Although MOTA is lower for RNN models, they have the lowest mostly lost counts with a low number of misses and IDS. It can be concluded that temporal modeling helps make a more robust model for occlusion in baseline ResNet50 models. However, when we apply TA to CBAM, despite increasing the overall metrics for modeling with ID loss, CBAM trained on ID + triplet loss struggles slightly. A likely reason is that the triplet loss is known to be difficult to converge, and the last epoch is overfitting to the dataset. We can interpret it as that temporal modeling improves partially tracked and mostly lost values.

How do State-of-the-Art Models Stack Up? The purpose of the Re-ID model in the tracking-by-detection framework is to reduce the number of IDS and loss of tracking. In this matter, RGA-SC makes tremendous improvements over the baseline methods with the comparatively lowest IDS value of 2785. PCB+RPP (I+T) has the second lowest IDS value with 3526. CBAM (I+T) and OSNet (I+T) have the highest and second-highest MOTA values. We believe that state-of-the-art Re-ID models prove useful in creating a more robust Re-ID model for MOT.

5.2 Benchmarking on MOT20

In the previous section, we analyzed the effects of various Re-ID models and training methods on sequence 05. We have discovered several key aspects of these results:

- Most models conditioned on sequential data using video-based Re-ID could increase MOTA and reduce IDS
- Using batch hard triplet loss along with identity loss also tends to increase MOTA and reduce IDS
- Recent state-of-the-art Re-ID models show promise in further reducing IDS

Since there is a limitation in submitting trackers to the benchmark server, we choose four methods, including the baseline method for benchmarking on MOT20. We show these Re-ID methods applied on Deep SORT and Tracktor++ (Bergmann et al., 2019), one of the state-of-the-art methods for the MOTChallenge benchmark. The benchmark results are shown in Table 5 along with results of other online MOT methods. Metrics used in the results are the same as Table 4 with the addition of false alarm per frame (FAF).

Comparison with Baseline Re-ID. Generally, for DeepSORT tracker, state-of-the-art Re-ID models score higher in IDF1, lower FP, higher precision, and lower IDS. It is no surprise for these models to have lower IDS since it is in our best interest to include Re-ID models for this purpose, and it proves that recent Re-ID models provide an edge over the baseline. However, MOTA scores do not change but are lower by a small margin for PCB+RPP and RGA-SC. As we see in Table 4 in the RGA-SC with I+T, we believe this to be a trade-off of reducing the number of IDS.

Application to Recent Tracker. We have applied CBAM and RGA-SC to a recent tracker that scores fairly high on the MOT20 benchmark; namely, Tracktor++ (Bergmann et al., 2019). Tracktor++ follows a method similar to tracking-by-detection can use the Re-ID model for data association. As for our implementation of Tracktor++, we use the official implementation, and the tracker parameters for that is used for MOT17 benchmark since parameters for MOT20 are not available and could not recreate the results posted on MOT20 benchmark (Tracktor++v2 is the official results). Comparison of our Tracktor++ results shows that the only metric that changed is the number of IDS. CBAM reduces the number of IDS, but on the contrary, RGA-SC increases the number of IDS. Note that the Re-ID model weights are the same as the one used for DeepSORT. By default, Tracktor++ does not fully utilize the Re-ID models, and we would need further analysis on Tracktor++ to figure out parameters to change so that Re-ID models could influence the tracking results.

Further Analysis and Future Works. Overall, state-of-the-art methods, temporal modeling, and metric

Table 5: Benchmark results on MOT20 Dataset. Along with our various DeepSORT trackers, we analyze the effects of Re-ID models on Tracktor++ (Bergmann et al., 2019). For comparison, we add the benchmarking results of SORT (Bewley et al., 2016), Tracktor++v2 (Bergmann et al., 2019), and Fair (Zhan et al., 2020). Note that we could not recreate the results of Tracktor++v2 on MOT20 since the hyperparameters for the tracker are not available (*Re-ID model for Tracktor++v2 is ResNet50 trained on MOT17 dataset with ID loss and batch hard triplet loss. **Re-ID model for Fair is ResNet34 and is jointly trained on the task of detection).

Tracker	Re-ID	MOTA ↑	IDF1 ↑	MOTP ↓	MT ↑	ML ↓	FP ↓	FN ↓	Rec. ↑	Prec. ↑	FAF ↓	IDS ↓	Frag. ↓
Deep SORT	BL	43.8%	43.7%	77.7%	212	311	31771	255254	50.7%	89.2%	7.1	3918	12048
	PCB+RPP	43.2%	46.0%	77.8%	203	318	30431	259659	49.8%	89.4%	6.8	3909	11396
	CBAM	43.8%	45.7%	77.7%	212	313	31658	255415	50.6%	89.2%	7.1	3677	11977
	RGA-SC	43.5%	47.3%	77.8%	208	319	29685	259403	49.9%	89.7%	6.6	3347	11239
Tracktor++	BL	52.0%	44.1%	78.7%	363	326	6988	236380	54.3%	97.6%	1.6	4813	4345
	CBAM	52.0%	44.1%	78.7%	363	326	6988	236380	54.3%	97.6%	1.6	4805	4345
	RGA-SC	52.0%	44.1%	78.7%	363	326	6988	236380	54.3%	97.6%	1.6	4814	4345
SORT	None	42.7%	45.1%	78.5%	208	326	27521	264694	48.8%	90.2%	6.1	4470	17798
Tracktor++v2	*	52.6%	52.7%	79.9%	365	331	6930	236680	54.3%	97.6%	1.5	1648	4374
Fair	**	61.8%	67.3%	78.6%	855	94	103440	88901	82.8%	80.6%	23.1	5243	7874

learning improve the Re-ID capability in the tracking-by-detection framework. It is well known that a big limitation of the tracking-by-detection framework is its heavy dependency on the detector. Fair (Zhan et al., 2020) scores the highest MOTA, the highest IDF1, the highest number of MT, and the lowest number of ML, which can be attributed to the joint learning of detection and embedding feature for Re-ID. As shown in Table 3, if we use a perfect detector, it is clear that there are minimal trade-offs between each of the metrics and we can see a relatively small gain on performance compared to the baseline method. Generally, there are no major differences in performance between each of the Re-ID models and makes us think that Re-ID has no effects on MOT metrics. However, as shown in Table 4 and 5, we can clearly see the performance gains by using state-of-the-art Re-ID methods such as IDF1 score, precision, IDS, and the number of fragmentation. It is future work to further understand the relationship between the Re-ID module and the detector, and provide a better way of data association in MOT.

For all our results, we use only the training sequences of MOT20 for training the Re-ID models. However, the original DeepSORT paper used the Mars dataset to train its Re-ID model (Wojke et al., 2017; Wojke and Bewley, 2018). We believe that using the various image and video Re-ID datasets will enhance the Re-ID model's capability since our Re-ID models are possibly overfitting to the limited scenes of MOT20.

6 CONCLUSIONS

In this work, we have improved the Re-ID model in the tracking-by-detection framework for online MOT. We use state-of-the-art Re-ID models along with various Re-ID modeling techniques such as metric learning and temporal modeling to improve the Re-ID metrics. Benchmarking on a heavily occluded MOT20 dataset proves that these methods improve identity switching and other metrics compared to the baseline models frequently used in this framework. We believe that the backbones used in the state-of-the-art Re-ID methods could be used with the recent joint detection and embedding (JDE) methods (Wang et al., 2019; Zhan et al., 2020) that score very high on the MOTChallenge, which will be one of the future work of our research.

ACKNOWLEDGEMENTS

This paper is partly based on results obtained from a project, JPNP16007, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This work was partly supported by JSPS KAKENHI Grant Number JP20J2212.

REFERENCES

Bergmann, P., Meinhardt, T., and Leal-Taixe, L. (2019). Tracking without bells and whistles. In *Proceedings of the IEEE international conference on computer vision*, pages 941–951.

Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468.

Choi, W. (2015). Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 3029–3037.

Dendorfer, P., Ošep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., and Leal-Taixé, L. (2020a). Motchallenge: A benchmark for single-camera multiple target tracking. *arXiv preprint arXiv:2010.07548*.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., and Leal-Taixé, L. (2020b). Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv:2003.09003[cs]*. arXiv: 2003.09003.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Fang, K., Xiang, Y., Li, X., and Savarese, S. (2018). Recurrent autoregressive networks for online multi-object tracking. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 466–475. IEEE.

Farhadi, J. R. A. and Redmon, J. (2018). Yolov3: An incremental improvement. *Retrieved September*, 17:2018.

Gao, J. and Nevatia, R. (2018). Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Leal-Taixé, L., Milan, A., Reid, I., Roth, S., and Schindler, K. (2015). MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*. arXiv: 1504.01942.

Liu, Y., Yan, J., and Ouyang, W. (2017). Quality aware network for set to set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5790–5799.

Luo, H., Gu, Y., Liao, X., Lai, S., and Jiang, W. (2019). Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Mahmoudi, N., Ahadi, S. M., and Rahmati, M. (2019). Multi-target tracking using cnn-based features: Cnnmtt. *Multimedia Tools and Applications*, 78(6):7077–7096.

McLaughlin, N., Del Rincon, J. M., and Miller, P. (2016). Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334.

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K. (2016). MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*. arXiv: 1603.00831.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., and Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496.

Wang, Z., Zheng, L., Liu, Y., Li, Y., and Wang, S. (2019). Towards real-time multi-object tracking. *arXiv preprint arXiv:1909.12605*.

Welch, G., Bishop, G., et al. (1995). An introduction to the kalman filter.

Wojke, N. and Bewley, A. (2018). Deep cosine metric learning for person re-identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 748–756. IEEE.

Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE.

Woo, S., Park, J., Lee, J.-Y., and So Kweon, I. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., and Yan, J. (2016). Poi: Multiple object tracking with high performance detection and appearance feature. In *European Conference on Computer Vision (ECCV)*, pages 36–42. Springer.

Zhan, Y., Wang, C., Wang, X., Zeng, W., and Liu, W. (2020). A simple baseline for multi-object tracking. *arXiv preprint arXiv:2004.01888*.

Zhang, Z., Lan, C., Zeng, W., Jin, X., and Chen, Z. (2020). Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3186–3195.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., and Tian, Q. (2016). Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 1116–1124.

Zheng, Z., Zheng, L., and Yang, Y. (2017a). A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):1–20.

Zheng, Z., Zheng, L., and Yang, Y. (2017b). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762.

Zhou, K., Yang, Y., Cavallaro, A., and Xiang, T. (2019). Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3702–3712.

Zhou, Z., Xing, J., Zhang, M., and Hu, W. (2018). Online multi-target tracking with tensor-based high-order graph matching. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1809–1814. IEEE.