VIP-HTD : A Public Benchmark for Multi-Player Tracking in Ice Hockey

Harish Prakash[1*] **Yuhao Chen**[2] **Sirisha Rambhatla**[3] **David Clausi**[4] **John Zelek**[5]
[1,2,3,4,5] Vision and Image Processing Group, System Design Engineering, University of Waterloo
{harish.prakash, yuhao.chen1, sirisha.rambhatla, dclausi, jzelek}@uwaterloo.ca

Abstract

Multi-Object Tracking (MOT) is the combined task of localization and association of subjects across a temporal sequence. Unlike the popular pedestrian tracking paradigms, monocular tracking of ice hockey players from broadcast feeds presents a variety of challenges due to rapid non-linear motions, occlusions, blurs, and pan-tilt-zoom effects. To tackle these issues, there neither exists public datasets nor benchmarks trained on public datasets to date. To this end, we propose: (a) VIP-HTD - a public ice hockey tracking dataset, processed & curated from existing work, and (b) a public benchmark for multi-player tracking based on it. Further, we also present our observations processing this dataset and discuss the two key metrics (IDF1 score and ID switches) required for optimal tracking evaluations. With this work, we take a step towards creating a unified public benchmark for evaluating multi-player tracking in hockey. Our dataset is available at `https://github.com/harshap-ai/VIP-HTD`

1 Introduction

Multi-Object Tracking and Identification pursues the combined tasks of detection, association, and re-identification across a temporal sequence of frames. Most state-of-the-art methods [1–3] leverage the Tracking-by-Detection (TBD) paradigm to predominantly track crowded pedestrian scenes and benchmark on the MOT Challenge Dataset [4]. Several methods [5–10] have extended this to tracking players in sports, either through handcrafted features or via deep learning methods.

Although tracking in sports is filled with practical applications, it has several inherent challenges in the form of non-linear motion, occlusion, dynamic actions, pan-tilt-zoom changes in the telecast, fragmented broadcast videos (commercials, pop-ups, crowd-cam), etc. Ice hockey, being one of the fastest field games, has all the aforementioned challenges and therefore, serves as an excellent candidate sport for overcoming these difficulties. One of the primary goals in ice hockey is to estimate the most strategic positions to place players on ice, such that, they are optimized to score goals and avoid conceding to the opposition [11]. To achieve this, the integral first step is the localization of players on the rink and tracking their movements consistently throughout a temporal sequence.

Recent approaches using deep learning have shown a multi-fold increase in the accuracy of tracking players compared to traditional methods [12, 13]. The existing state-of-the-art (SOTA) baseline in ice-hockey [8] uses an MOT Neural Solver (MPN) [14] architecture fine-tuned on a private dataset with broadcast ice-hockey clips to track players with high accuracy (MOT accuracy score). Subsequently, they [9] use this framework to generate tracklets (tracks) for the downstream task of player identification from broadcast videos. While these are promising developments, there exist two limitations in the baseline: first, they do not benchmark on a public dataset, making it difficult to reproduce their results; second, the MOT accuracy favors the detector more than the tracker itself. Though one can argue that in TBD paradigms, detectors play an important role along with the Trackers, it is essential to evaluate them separately to know where the bottleneck lies.

To address the first limitation, as far as we know, there exists only a single public dataset with MOT annotations for Ice-hockey [15]. Their contribution is salient; but, the dataset remains unusable - it has erroneous/missing annotations, redundant clips, wrong meta-data, and no parsed frames (Ref. *Figure* 1). Thus, to make it plug&play, we take up the task of processing this dataset extensively and build a new ready-to-use version titled **"VIP-Hockey Tracking Dataset (VIP-HTD)"** based on it. Taking a step forward, we use our dataset to create a public benchmark using a simple graphical Message Passing Network (MPN) to enable future performance evaluations.

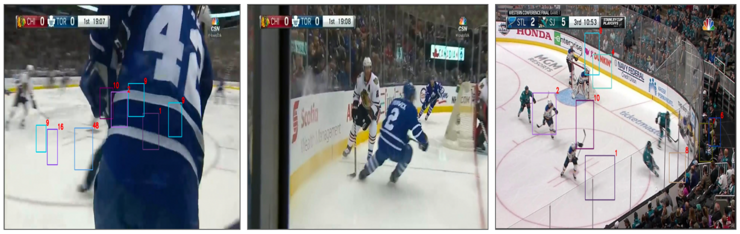

Fig. 1: MHPTD failure modes: **L-R** {Erroneous, Missing & Offset} bounding boxes,

To address the second limitation, we observe that the IDF1 and ID switch (IDs) scores are the principle tracking metrics that establish whether a player was consistently tracked throughout a temporal sequence. Recognizing its importance in reducing player swaps/tracklet fragmentations, we model our benchmark based on optimizing these metrics. Our contributions in this work are three-fold:

1. We propose a plug&play dataset (VIP-HTD), processed and curated from the MHPT [15] dataset.
2. We train a similar model as [8, 9, 14] on our proposed dataset to create benchmark scores, and,
3. We discuss the relevance of IDF1 and ID Switch scores [16] in generating tracklets without swaps/fragmentations.

2 Dataset

The motivation behind the VIP-HTD dataset is to establish a public ice hockey-specific tracking dataset. Instead of manually annotating frames, we improved upon the raw MHPT dataset [15], through the addition of parsed frames, rectification of erroneous annotations, curation of non-redundant feed, and proper meta-data analysis, making it ready-to-use. Our derived VIP-HTD consists of 22 broadcast clips from 8 different games, containing side-of-the-rink views with a resolution of 1280x720p.

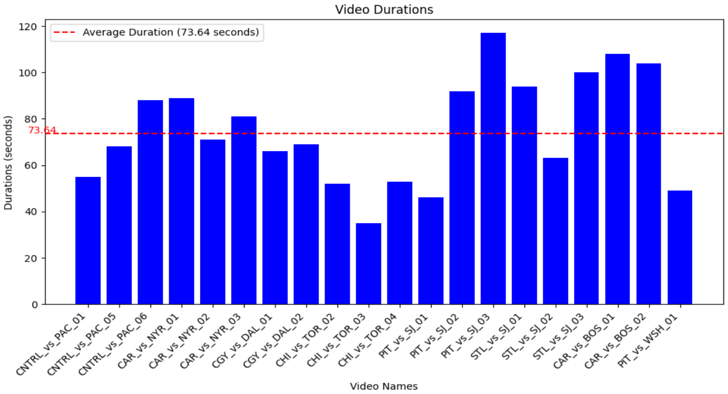

Fig. 2: Duration of each video in the VIP-HTD dataset.

The average duration of each clip is 73 seconds, with frames sampled at either 30Hz or 60Hz. The refined annotations contain *{frame IDs, objects IDs, bounding box coordinates, confidence score, category, visibility}*, with the confidence score and category set to 1, as they are manual annotations and one single category (players), respectively. The preferred train:test:validation split that works best

[*1] https://github.com/opencv/cvat

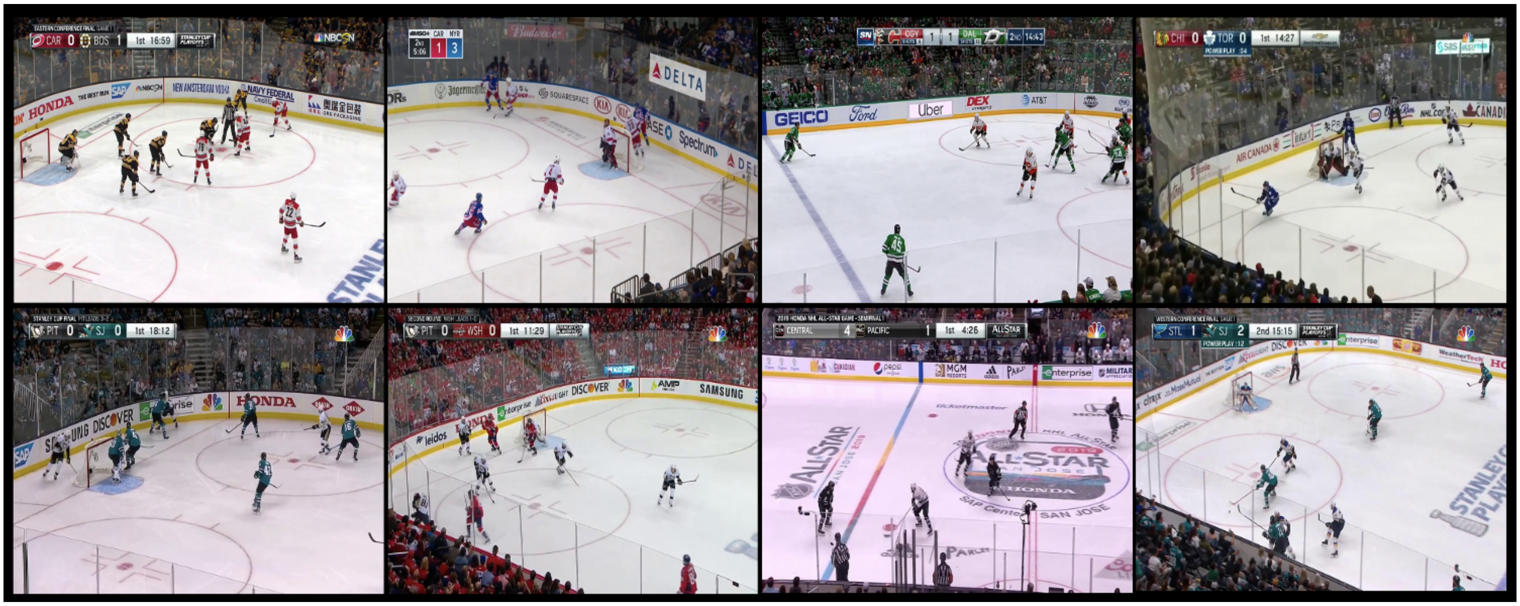

Fig. 3: Frame samples from the 8 different games in VIP-HTD dataset

in our benchmark experiments is 14:7:1, with the test set belonging to similar games from the train set, but containing mutually exclusive frames (different clips/portions) to avoid data leakage.

Two major observations & contributions made during the creation of our VIP-HTD dataset are as follows:

2.1 Frames

We parsed all the clips according to their respective sampling rates using the open-source CVAT tool[1]. Interestingly, when we tried parsing frames using external libraries like OpenCV [17] or moviepy [18], we ended up with more frames compared to the frames exported from CVAT for a given clip. Upon further investigations, we found that the cause of this deviation was different video synchronization options used in different tools - specifically, CVAT uses FFMPEG [19] as a backend API for parsing videos, which passes timestamps from the Demultiplexer to the Multiplexer *(enable flag: -vsync 0)*, facilitating synchronized frame parsing as per the clip's native sampling rate. Without this option, we observed an offset effect which translates the player bounding box location by a small margin, as we move toward the tail-end of the parsed frames distribution. Therefore, to avoid this error, we have provided the synchronized frames extracted from CVAT along with this dataset.

2.2 Annotation Scheme

Conventionally, in the MOT challenge datasets [4], an object (pedestrian) is assigned identities at the 'tracklet' level i.e., whenever an object exits at time, t_i and re-enters at time t_j s.t $j > i$, it is assigned a new identity. Conversely, in both the raw MHPTD [15] and Kanav et al.'s [8] private dataset, object (player) identities are assigned at the 'personnel' level i.e., when a player exits at time, t_i and re-enters at time t_j s.t $j > i$, they are assigned the same identity. The reasoning behind it is that pedestrians often move unidirectionally while hockey players appear repetitively in a given field of view (FoV).

Although this intuitively makes sense, without an ideal re-identification model and a large temporal window, a player who re-enters after a significant time interval is often assigned a new object ID. Moreover, re-identification is especially difficult in ice hockey, since player features are often indistinguishable due to covered faces, uniform jersey colors (within a team), and limited appearance features.

During evaluation, when ground-truth tracklets T^{gt}_{n*} with time-ordered object IDs O^{id}_n, are matched to tracked hypotheses ID h^{id}_m based on a distance threshold (IoU), we obtain tracked outputs as $T^{det}_{n*} = (h^{id}_m, o^{id}_n)$, where m and n are tracked and ground truth object IDs respectively. An object re-entering after a frame interval $t > 1$, which is assigned a new h^{id}_k leads to a new (h^{id}_k, O^{id}_n) pair $(h_k id, o^{id}_n)$, s.t $k \neq m$. This instant is then accounted as an identity switch (IDs), resulting in an ↑ in IDs and ↓ in IDF1 score. All further associations are then updated as $(h_k id, o^{id}_n)$, and the new hypothesis $h_k id$ is considered as the default (until another switch occurs).

The 'personnel-level' annotation scheme, thus penalizes models without a strong re-identification component for re-entering subjects, which is unsuitable for certain tracking models without a large temporal window [14, 20, 21]. Thus, it is essential to process the ground-truth annotations to have both 'tracklet-level' (as followed in [4]) and 'personnel-level' IDs, for use as per the chosen model's abilities. We provide both the annotation schemes, along with their conversion script in the code base.

3 Methodology

3.1 Benchmark Implementation

We opt for the Message Passing Network architecture (MPN)[14] similar to our baseline [9] to ensure a fair comparison, but our implementation benchmarks on the public VIP-HTD dataset while the baseline is trained using a private, unpublished hockey dataset. The MPN tracker encompasses three major components: A Detector, A Re-ID feature extractor, and an MPN tracker. Player detection is performed using the Tracktor [20] algorithm, which uses a Faster-RCNN network [22] with a ResNet-50 [23] based Feature Pyramid Network (FPN) [24] backbone, pre-trained on the COCO dataset [25] and fine-tuned on our VIP-HTD dataset. Since Tracktor is originally a tracking algorithm, it generates its own object IDs, which are discarded. The appearance features of these detections are encoded using a Resnet-50 [23], pre-trained on the ImageNet [26] dataset, followed by global average pooling and two fully-connected layers to obtain embeddings of dimension 256. This network is fine-tuned by [14] on three popular Re-Identification (ReID) datasets: Market1501 [27], CUHK03 [28] and DukeMTMC [29]. Note that we do not train the Re-ID network on our Hockey dataset and consider the original Re-ID network as used in [8, 9].

To train the MPN [14], we use a batch size of 8, where each batch corresponds to small clips of 15 frames uniformly sampled at 9 Hz. This approach helps handle the two different sampling rates (30 and 60 Hz) in the VIP-HTD dataset. The network is optimized for 25 epochs with AdamW [30], with a learning rate of 0.001 and weight decay of 0.0001. Data augmentation is performed by randomly removing nodes from the graph, thereby simulating missed detections, and randomly shifting bounding boxes. Both, the object detector [20], which is fine-tuned for 27 epochs using the default configuration; and

Method	FP↓	FN ↓	IDSW ↓	IDF1 ↑	MOTA ↑
Tracktor[20] (30Hz)	1706	4216	687	0.56	90.1%
Kanav et al. [8] (30Hz)	4057	2586	414	0.62	**94.0**%
Ours (30&60Hz)	14583	5627	**403**	**0.74**	82.0%

Table 1: Comparison between our model and baselines

Method	Annotation Scheme	False Positives ↓	False Negatives ↓	IDs ↓	IDF1 ↑	MOTA ↑
Kanav et al. [8]	Personnel-level	3838	2735	428	0.62	94.0%
Kanav et al. [8]	Tracklet-level	7865	5357	**251**	**0.77**	89.0%
Ours	Personnel-level	14452	5703	509	0.63	81.7%
Ours*	Tracklet-level	14583	5627	**403**	**0.75**	81.7%

Table 2: Comparison between 'Personnel-level' and 'Tracklet-level' annotation schemes

the MPN tracking algorithm [14] are trained on an NVIDIA GeForce 2080Ti GPU with 32 GB RAM and 16 cores. It is to be noted here that for [8], we fine-tune the object detector [20] and train the MPN network to recreate their results solely using their private dataset. For our baseline, we fine-tune the object detector [20] and train our MPN network solely with our VIP-HTD public dataset.

3.2 Evaluation Metrics

To quantitatively evaluate our results, we calculate the Multi-Object Tracking Accuracy (MOTA) [16] score and Identification F1 (IDF1) score as our metrics, as followed by most other benchmarks in the tracking space.

- The MOTA is estimated as the complement of three distinct errors -

$$MOTA = 1 - \frac{\sum_t FN_t + FP_t + IDs_t)}{\sum_t GT_t} \tag{1}$$

  - **False Positives** (FPs) occur when a hypothesis h_i has no corresponding ground truth gt_i within a specified threshold. This is given as:

$$\overline{FP} = \frac{\sum_t FP_t}{\sum_t GT_t} \tag{2}$$

  - **False Negatives** (FNs) occur when a ground-truth gt_i is missed to be detected, and has no corresponding hypothesis h_i, given by:

$$\overline{FN} = \frac{\sum_t FN_t}{\sum_t GT_t} \tag{3}$$

  - **ID switches** (IDs) occur when two different hypotheses h_1o_1 and h_2o_2 switch (frame-level) when in close proximity, and lead to h_1o_2 and h_2o_1, as given by:

$$\overline{IDs} = \frac{\sum_t IDs_t}{\sum_t GT_t} \tag{4}$$

    where, h and o are hypothesis and ground-truth IDs respectively.

- The **IDF1 score** helps estimate the percentage of correctly tracked identities (tracklet-level), given as the ratio of correctly identified detections over the average number of ground truth and computed detections.

$$IDF1 = 2 \times \frac{TP_{id}}{TP_{id} + FP_{id} + FN_{id}} \tag{5}$$

  where, $TP_{id}, FP_{id}, FN_{id}$ are True Positive, False Positive and False Negative tracklet identities.

3.3 What is relevant?

The False-positives (FPs) and False-negatives (FNs) are majorly dependent on the accuracy of the object detector and are not affected by the tracking algorithm. Thereby, the MOTA score is biased in favor of optimizing the detector rather than the tracker. This paints a false picture of our tracker's performance and leads us to a proxy objective.

To offset this effect, the two most important metrics to follow are: the IDF1 score, which measures how consistently the identity of a tracked object is preserved with respect to the ground truth identity, and thus is favored to ↑; and the IDs score, which increases when a ground truth ID i is assigned a hypothesis ID j, when the last known assignment ID was $k \neq j$, and thus is favored to ↓. A high IDF1 score and low IDs denote that our model has fewer tracklet fragmentation errors and player identity swaps, and is imperative for downstream analyses.

4 Results and Ablations

We compare our method trained on the VIP-HTD dataset [15], with the present baseline from Kanav et al. [8] and Tracktor [20], both of which were trained on the private hockey dataset used by [8, 9]. To ensure a fair comparison, we shrink our dataset's clip sizes to ∼40 seconds, similar to the clip durations of the private baseline dataset, as FPs and FNs tend to increase proportionally with the number of frames. Note that, we end up obtaining twice the frame count for 60 Hz clips when compared to 30 Hz clips when shrunk to the same duration, thereby cumulatively leading to more FPs and FNs in our results. From Table *1*, we find good improvements in the IDF1 and IDs scores, which are key factors for sports analytics as the identities of players remain relatively consistent throughout the sequence in our approach.

We perform an ablation study on the proposed change in the annotation scheme to better understand our approach. We perform experiments on both the private baseline dataset [8, 9] and our VIP-HTD dataset for a fair comparison. In Table *2*, for both datasets, we achieve significantly higher IDF1 scores and lower IDs for 'Tracklet-level' annotations. This is because the MOT neural solver MPN architecture [14], that both we and the baseline [8, 9] adopt, has a temporal window of only 15 frames. Thus, a player who re-enters after 15 frames will be given a new ID, which leads to an ID switch error in the 'Personnel-level' annotation scheme. It is also interesting to note here that while IDF1 scores ↑ and ID switches ↓ significantly, the MOTA score remains constant (nearly) for both cases, thereby re-affirming that it is detector-dependent and doesn't depict our tracking performance directly.

5 Conclusion & Future Works

We present a public, plug&play dataset (VIP-HTD) for the challenging task of Ice-hockey player tracking. We explain the processes undertaken to process and curate it and present a study on two different annotation schemes that can be followed based on the model design. We disseminate the differences between different evaluation metrics and establish IDF1 score and IDs as the two major metrics to consider for sports tracking. Going forward, we aim to explore different architectures with the VIP-HTD dataset, investigate the drop in MOTA further, and progress towards achieving downstream tasks that are aided by this method.

Acknowledgments

This work was supported by Stathletes through the Mitacs Accelerate Program and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "Bot-sort: Robust associations multi-pedestrian tracking," 2022.

[2] Z. Liu, X. Wang, C. Wang, W. Liu, and X. Bai, "Sparsetrack: Multi-object tracking by performing scene decomposition based on pseudo-depth," 2023.

[3] Y.-H. Wang, J.-W. Hsieh, P.-Y. Chen, M.-C. Chang, H. H. So, and X. Li, "Smiletrack: Similarity learning for occlusion-aware multiple object tracking," 2023.

[4] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv:1603.00831 [cs]*, Mar. 2016, arXiv: 1603.00831. [Online]. Available: http://arxiv.org/abs/1603.00831

[5] K. Okuma, A. Taleghani, D. Freitas, J. Little, and D. Lowe, "A boosted particle filter: Multitarget detection and tracking," vol. 3021, 05 2004.

[6] T. Misu, M. Naemura, W. Zheng, Y. Izumi, and K. Fukui, "Robust tracking of soccer players based on data fusion," in *2002 International Conference on Pattern Recognition*, vol. 1, 2002, pp. 556–561 vol.1.

[7] Y. Cai, N. Freitas, and J. Little, "Robust visual tracking for multiple targets," 05 2006, pp. 107–118.

[8] K. Vats, M. Fani, D. A. Clausi, and J. S. Zelek, "Evaluating deep tracking models for player tracking in broadcast ice hockey video," 2022.

[9] K. Vats, P. Walters, M. Fani, D. A. Clausi, and J. S. Zelek, "Player tracking and identification in ice hockey," *CoRR*, vol. abs/2110.03090, 2021. [Online]. Available: https://arxiv.org/abs/2110.03090

[10] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy, "Learning to track and identify players from broadcast sports videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1704–1716, 2013.

[11] A. Thomas, "The impact of puck possession and location on ice hockey strategy," *Journal of Quantitative Analysis in Sports*, vol. 2, pp. 6–6, 02 2007.

[12] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

[13] H. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistic Quarterly*, vol. 2, 05 2012.

[14] G. Brasó and L. Leal-Taixé, "Learning a neural solver for multiple object tracking," *CoRR*, vol. abs/1912.07515, 2019. [Online]. Available: http://arxiv.org/abs/1912.07515

[15] K. C. Yingnan Zhao, Zihui Li, "A method for tracking hockey players by exploiting multiple detections and omni-scale appearance features," *Project Report*, 2020.

[16] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, 01 2008.

[17] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[18] Z. Teed and S. M. Langnickel, "Moviepy: A python library for video editing," https://zulko.github.io/moviepy/, 2018.

[19] FFmpeg Developers, "Ffmpeg," https://www.ffmpeg.org/, 2000.

[20] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. [Online]. Available: https://doi.org/10.1109%2Ficcv.2019.00103

[21] Y. Zhang, P. Sun, Y. Jiang, D. Yu, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," *CoRR*, vol. abs/2110.06864, 2021. [Online]. Available: https://arxiv.org/abs/2110.06864

[22] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[24] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *CoRR*, vol. abs/1612.03144, 2016. [Online]. Available: http://arxiv.org/abs/1612.03144

[25] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2015.

[27] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124.

[28] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[29] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," vol. 9914, 10 2016.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7