Extending IOU Based Multi-Object Tracking by Visual Information

Erik Bochinski, Tobias Senst and Thomas Sikora
Communication System Group, Technische Universität Berlin
Einsteinufer 17, 10587 Berlin
{bochinski, senst, sikora}@nue.tu-berlin.de

Abstract

Today's multi-object tracking approaches benefit greatly from nearly perfect object detections when following the popular tracking-by-detection scheme. This allows for extremely simple but accurate tracking methods which completely rely on the input detections as the high-speed IOU tracker. For real world applications, few missing detections cause a high number of ID switches and fragmentations which degrades the quality of the tracks significantly. We show that this problem can be efficiently overcome if the tracker falls back to visual single-object tracking in cases where no object detection is available. In several experiments we show for different visual trackers that the number of ID switches and fragmentations can be reduced by a large amount while maintaining high tracking speeds and outperforming the state-of-the art for the UA-DETRAC and VisDrone datasets.

1. Introduction

The most successful multi-object tracking (MOT) approaches follow the tracking-by-detection paradigm. Typical methods use sophisticated appearance models to re-identify object instances over a long temporal range [25, 4] or perfrom complex global optimizations [2, 29, 3] to compute the tracks for each object.

However, in recent years object detectors have been improved significantly. The underlying advancements are not at least fueled by the current deep-learning era [16, 13] and large-scale object detection benchmarks [18, 8]. This results in highly accurate object detection methods like Fast(er)/Mask R-CNN [11, 24, 12], FCN [20] or SSD [19].

This poses a shift in the requirements for tracking algorithms. Being able to rely on these increasingly accurate object detections allows for much simpler tracking-by-detection approaches [6, 5, 30]. They all share the core principle of associating detections with an high spatio-temporal overlap to individual tracks. This overlap is measured by the intersection-over-union of the detection bounding boxes between consecutive frames.

The Simple Online Realtime Tracker (SORT) [5] employs a Kalman filter motion model and solves the assignment problem of the detections optimally using the Hungarian algorithm. This method was extended in [30] by integrating appearance information through a deep association metric to handle longer periods of occlusion. Similar to SORT, the intersection-over-union Tracker (IOU) [6] relies only on detections and does not use any image information. This simple tracker employs no motion model and associates the detections to tracks in a greedy manner. As a result the IOU tracker can operate at thousands of frames per second (assuming the required detections are available) while outperforming much more complex state-of-the-art methods [21]. A major drawback of this simple approach is its requirement for a high recall of the underlying detector. Every gap caused by a single or few missing detections leads not only to false negatives but also to the termination and restart of the track, causing high rates of fragmentation and ID switches.

In this work, we approach this problem by incorporating a visual single-object tracker into the IOU tracking scheme to increase the robustness against missing detections. Our idea is to continue each track by a visual tracker if no new detection can be associated and thus fill the gaps between tracks. This reduces the fragmentation of the tracks and number of ID switches. As a side-effect this allows to apply a higher detection confidence threshold as the method no longer relies on continuous detections over all frames. The accuracy and speed of the proposed approach depends primarily on the object detectors and visual single-object trackers performances. In our experiments on the UA-DETRAC [28] and VisDrone [31] datasets, we investigate the influence of different state-of-the-art multi-object detectors and single-object trackers to the performance of our framework. We show that this approach leads to an improved overall accuracy and effectively reduces the amount of ID switches and fragmentations by a great portion while maintaining a low computational footprint and outperforms the state-of-the art for both datasets.

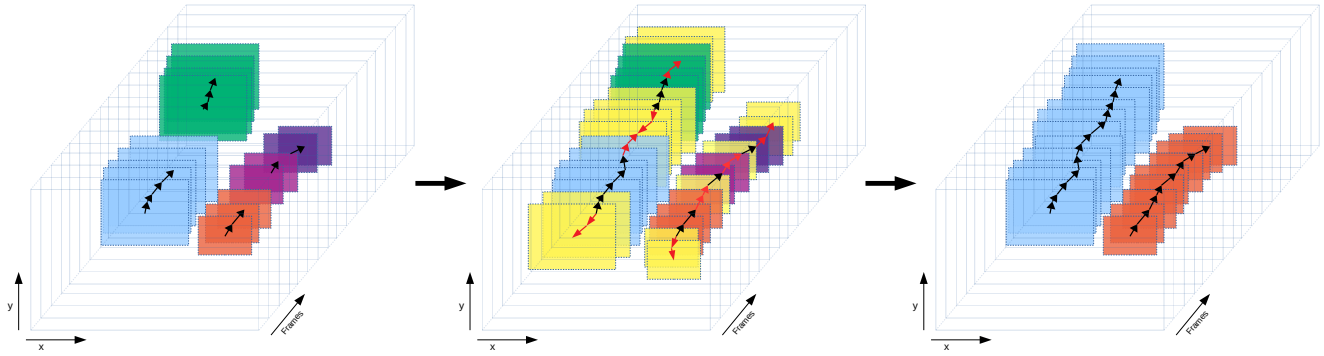


Figure 1. Basic principle of the extension for the IOU Tracker: IOU tracks are usually highly fragmented due to missing detections (left). Gaps can be filled by employing a visual tracker to compensate for missing object locations (yellow, middle). The resulting tracks are less fragmented (right).

2. Method

The visual intersection-over-union (V-IOU) tracker is a conceptually consistent continuation of the IOU tracking approach presented in [6]. It is designed to reduce the high amount of ID switches and the high rate of fragmentation of the resulting tracks of the baseline method. This is achieved by incorporating a visual single-object tracker into the tracking framework to compensate for missing object detections.

IOU Tracker We will first briefly review the concept and limitations of the baseline IOU Tracker. The main idea behind this approach is that current object detectors are reliable enough so that failures induced by false positive/negative detections can be ignored. Based on this assumption, the task of multi-object tracking following the tracking-by-detection paradigm becomes trivial.

The IOU tracker associates detections of subsequent frames solely by their spacial overlap to tracks. In the original proposed approach, this is done in a greedy way: A track gets the detection with the highest intersection-over-union to its last known object position (i.e. the previous detection of the track) assigned. Although not optimal, this heuristic approach can be sufficient. Alternatively, the association can be done by solving a linear assignment problem as in [5] using the Hungarian algorithm. In real-world applications, false positive/negative detections occur and will interfere with the tracking process. Therefore, the resulting tracks are filtered by requiring each track to contain at least one detection satisfying a high detection confidence ($\geq \sigma_h$) and to have a minimum length of at least t_{min} frames. This effectively sorts out many failure cases caused by false-positive detections. False negative detections on the other hand cause a track immediately to stop. The IOU tracker will not propagate the last detection, thus a new track will be created at the next available one. This leads to a high rate of ID switches and fragmentation of the tracks.

Visual Tracking Extension False negative detections pose a general problem to tracking-by-detection approaches, especially for the IOU tracker as missing detections are not propagated. We therefore propose to extend the IOU tracker by falling back to visual single-object tracking in the case where no detection is available for association. The full visual tracking extension is shown in Fig. 1.

Visual tracking is performed in two directions: First, if no detection satisfies the σ_{IOU} threshold of [6] for a track, the visual tracker is initialized on the last known object position (the detection at the previous frame) and used to track the object for a maximum of ttl frames. If a new detection satisfies the σ_{IOU} threshold within these ttl frames, visual tracking is stopped and the IOU tracker is continued. Otherwise, the track is terminated. This is usually sufficient to compensate reliably for few missing detections.

However, with increasing number of visual tracked frames it becomes more likely that the visual tracker looses the track or jumps over to another object. To limit the number of consecutive frames where the object is only tracked by visual cues, we perform the visual tracking also backwards through the last ttl frames for each new track. If the overlap criteria is met for an existing, finished track we merge them. In this way it is possible to close gaps of an length of up to $2 \cdot ttl$ frames whereas the single visual object trackers are employed only for a maximum of ttl frames.

Although the visual forward and backward tracking of the object helps to merge discontinued tracks it also adds visually tracked stubs at the beginning and end of each completed track as seen in Fig. 1(b). As tracks should start when the object enters the scene and end when the object left, these stubs can not contribute to merge gaps and are prone to contain errors as the object of interest may not be in the scene. For this reason we limit the usage of visual tracking for gap closure and cut these visually tracked stubs of bounding boxes off the tracks.

Tracker	Detector	σ_h	σ_{iou}	t_{min}	ttl
IOU	CompACT	0.2	0.4	2	-
	Mask R-CNN	0.95	0.6	7	-
V-IOU	CompACT	0.3	0.5	3	12
	Mask R-CNN	0.98	0.6	13	6

Table 1. Best parameters for all tracker/detector combinations for the UA-DETRAC dataset.

3. Experiments

We evaluate the proposed V-IOU tracker on two common multi-object tracking datasets. First we compare the performance of the IOU and V-IOU tracker on the UA-DETRAC [28] dataset and investigate their abilities to deal with with false negative detections. Additionally, we study the impact of different visual single-object trackers on the VisDrone [31] dataset. Finally, we provide a comparison of our method to the state-of-the-art for both datasets. The best tracker configuration for each experiment has been determined by grid search of the trackers parameters.

3.1. UA-DETRAC

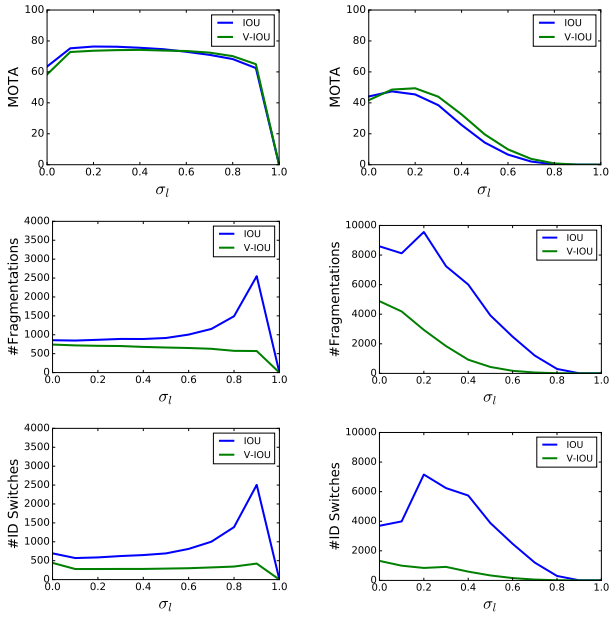
The UA-DETRAC dataset consists of 10 hours of fully annotated video material for vehicle tracking, split into 60 training and 40 test videos. It was recorded from typical traffic monitoring views and covers a wide variety of scale, pose, illumination, occlusion and weather conditions. The evaluation follows the *PR* based metrics presented in [28] which require the tracker to run well at all confidence levels of the detector.

Detections We perform the experiments using detections of two different methods: First, we evaluate the detections of CompACT [7] which are provided along with the dataset. This detector was trained on the UA-DETRAC train split where it achieves an *AP* of 77.37%. For the test set, the *AP* is 53.23%.

Secondly, we computed detections of Mask R-CNN [12] using a model trained on the COCO trainval35k dataset [18] which achieves a *mAP@IoU=0.50:0.95* of 46.5 on the COCO minival set. We used only detections of the car, bus and truck classes for our experiments which achieve an *AP* of 80.53% on the UA-DETRAC train set and 80.48% on the test set. The detections are made publicly available[1]. Beside the higher average precision compared to CompACT, the detectors performance is the same for both the UA-DETRAC training and testing data which is important since the trackers parameters can only be tuned on the training set.

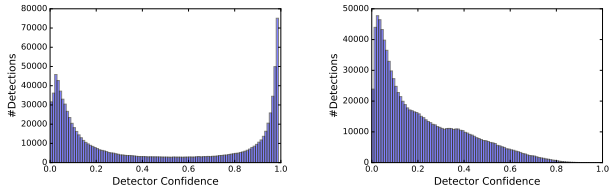
Tracking The baseline IOU and extended V-IOU trackers are optimized for the best PR-MOTA value on the UA-DETRAC train sequences for both previously described detectors. Only the fast KCF tracker [14] was considered for the evaluations of V-IOU. The parameters for the best performing configurations are shown in Tab. 1. Note that the parameter σ_l is varied from 0 to 1.0 for the *PR* based evaluation of the benchmark.

[1] https://github.com/bochinski/iou-tracker.git



(a) Mask R-CNN (b) CompACT

Figure 2. Comparison of the of the baseline IOU and our extended V-IOU trackers in terms of MOTA(↑), #Fragmentations(↓) and #ID Switches(↓) at different low detector thresholds σ_l for different detection methods. Supplementing the the IOU tracker with visual information produces much less fragmentations/ID switches and therefore more useful tracks.



(a) Mask R-CNN (b) CompACT

Figure 3. Distribution of confidence scores for the Mask R-CNN and CompACT detections on the UA-DETRAC test set.

Results The evaluation results for both the baseline IOU tracker and our extended V-IOU tracker using the KCF are compared to the state-of-the-art in Tab. 2. When using CompACT detections, the PR-MOTA is improved by an additional 1.6% (10% relative), PR-FM is reduced by a factor of 3 and PR-IDs by a factor of 6. For Mask R-CNN detections, the PR-MOTA value does not change from IOU to V-IOU but the number of PR-ID and PR-FM improves also by a factor of 4 and 2.5 respectively.

Method	Detector	PR-MOTA(↑)	PR-MOTP(↑)	PR-MT(↑)	PR-ML(↓)	PR-IDs(↓)	PR-FM(↓)	PR-FP(↓)	PR-FN(↓)	Speed(↑)
GMPHD-KCF [17]	CompACT	14.8%	36.0%	14.0%	**18.1%**	798.8	1606.8	38596.6	174042.7	24.60 fps
GMPHD [17]	CompACT	14.1%	36.3%	13.2%	19.0%	797.2	2143.8	38032.4	177215.1	45.24 fps
CEM$^{\#}$ [2]	CompACT	5.1%	35.2%	3.0%	35.3%	267.9	352.3	**12341.2**	260390.4	4.62 fps
CMOT$^{\#}$ [4]	CompACT	12.6%	36.1%	16.1%	18.6%	285.3	1516.8	57885.9	167110.8	3.79 fps
GOG$^{\#}$ [23]	CompACT	14.2%	37.0%	13.9%	19.9%	3334.6	3172.4	32092.9	180183.8	390.00 fps
DCT$^{\#}$ [3]	R-CNN	11.7%	38.0%	10.1%	22.8%	758.7	742.9	336561.2	210855.6	0.71 fps
H^{2}T$^{\#}$ [29]	CompACT	12.4%	35.7%	14.8%	19.4%	852.2	1117.2	51765.7	173899.8	3.02 fps
IHTLS$^{\#}$ [9]	CompACT	11.1%	36.8%	13.8%	19.9%	953.6	3556.9	53922.3	180422.3	19.79 fps
IOU [6]	R-CNN	16.0%	**38.3%**	13.8%	20.7%	5029.4	5795.7	22535.1	193041.9	100.84 fps
IOU [6]	EB	19.4%	28.9%	17.7%	18.4%	2311.3	2445.9	14796.5	171806.8	6.90 fps
IOU [6]	CompACT	16.1%	37.0%	14.8%	19.7%	2308.1	3250.4	24349.4	176752.8	**327660** fps
IOU [6]	Mask R-CNN	**30.7%**	37.0%	30.3%	21.5%	668.0	733.6	17370.3	179505.9	14956 fps
V-IOU	CompACT	17.7%	36.4%	17.4%	18.8%	363.8	1123.5	26413.3	**166571.7**	1117.90 fps
V-IOU	Mask R-CNN	**30.7%**	37.0%	**32.0%**	22.6%	**162.6**	**286.2**	18046.2	179191.2	359.18 fps

Table 2. Best results for each Tracker on the overall DETRAC-Test dataset including the easy, medium and hard splits. The results of $^{\#}$ were taken from [28].

The MOTA metric is influenced by the number of false positive and negative bounding boxes of the tracks and the number of ID switches. All three components are weighted equally. This means that the number of ID switches has only a minor impact on the overall MOTA scores. The better the input detections for the tracker, the less visually tracked bounding boxes need to be inserted by the V-IOU approach. Hence there is only a small improvement in the PR-MOTA score when using CompACT and Mask R-CNN detections but still a huge advance in terms of PR-IDs and PR-FM. This behaviour can be seen in Fig. 2 where the respective metrics are plotted over all detector thresholds as used for the PR based evaluation.

Figure 2(a) shows the results for Mask R-CNN with a good performance for all thresholds σ_l expect for the lower and upper bound of 0.0 and 1.0. The number of ID switches and fragmentations for the baseline IOU tracker increases considerably for higher detector thresholds as more less-confident detections are filtered out. Hence, the initial assumption of one detection per object and frame is violated with an increasing frequency. Our extended V-IOU tracker on the other hand is able to compensate for these missing detections reliably. Although this affects the MOTA value only marginally, the subjective quality of the tracks improves considerably for most applications.

A similar can be seen in Fig. 2(b) when using the CompACT detections. The position of the peak MOTA and worst number of ID switches/fragmentations is caused by different distributions of the confidence scores for each detector as shown in Fig. 3. As the confidence threshold is increased, more detections are removed earlier compared to Mask R-CNN which can be compensated by the visual tracking of V-IOU. This is also noticeable in the MOTA score for $\sigma_{IOU} > 0.2$.

In general, the evaluation on the UA-DETRAC dataset shows that the tracking performance is considerably improved by the visual tracking extension of the IOU tracker. Since the tracker uses, in contrast to the baseline method, image information, the runtime is decreased. With over 1,000 and 350 fps using the CompACT and Mask R-CNN detections on average over all confidence thresholds, the method can still be considered to be high speed.

3.2. VisDrone

The VisDrone dataset was captured in urban and country environments using different kinds of video drones. The evaluation benchmark targets object detection in videos and images, single-object tracking and multi-object tracking.

The sequences for multi-object tracking are split into 56 training, 7 validation and 16 test videos. Detections of Faster R-CNN [24] are supplied. In contrast to UA-DETRAC, the VisDrone MOT dataset contains 4 different categories of objects to track: car, bus, truck, van and pedestrian. The IOU tracking approach is originally not designed to handle different object classes simultaneously. We therefore neglect the classes while tracking and assign the class with the largest number of detections to the track after it is finished. Similar, we perform non-maximum suppression with an overlap threshold nms_t ignoring the class labels and use only the highest confidence scores.

Since the detector was trained on the training split of the dataset we decided to tune the parameters for the validation split. Two different visual single-object trackers where evaluated: KCF [14] as for UA-DETRAC and Medianflow [15]. The parameter tuning is done for the baseline IOU and the extended V-IOU tracker with two different objectives: First, we determine the parameters for the best overall performance. Secondly, we optimize the parameters for the best average MOTA over all sequences to have an equal weighting of the scenes with different lengths and numbers of objects. Only for the second strategy we exclude sequence `uav0000268_05773_v` since it contains a large amount of false positive detections, probably caused by overexposure of the recording. The results are presented in Tab. 3. where the first rows shows the best configuration and performance regarding the overall MOTA and the second rows the best average MOTA for each tracker. Instead of assigning the detections to tracks in a greedy manner as described in

Configuration							Results on Validation Set										
Tracker	nms_t	σ_l	σ_h	σ_{iou}	t_{min}	ttl	IDF1(↑)	IDP(↑)	IDR(↑)	MT(↑)	ML(↓)	IDs(↓)	FM(↓)	FP(↓)	FN(↓)	MOTA (AVG)(↑)	MOTP(↑)
-	0.6	0.5	0.98	0.05	7	-	40.9	68.5	29.1	102	297	177	435	5736	46979	26.4	78.1
	0.6	0.5	0.95	0.05	15	-	41.2	65.0	30.2	102	288	177	488	7329	45768	25.9 (31.3)	77.6
KCF	0.6	0.9	0.98	0.05	23	8	45.3	75.6	32.4	105	304	75	387	5592	46532	27.3	77.8
	0.6	0.7	0.95	0.2	19	4	44.2	64.7	33.6	110	279	136	597	9109	43679	26.3 (32.3)	76.8
Medianflow	0.6	0.9	0.98	0.1	23	8	45.7	75.8	32.7	107	304	70	387	5523	46390	27.6	77.8
	0.5	0.7	0.98	0.2	19	10	46.5	69.7	34.9	117	291	78	524	8261	44233	26.8 (32.7)	77.1

Table 3. Comparison of the best configurations for the VisDrone validation set for the baseline IOU tracker and the proposed extension using the KCF and Medianflow visual single object trackers. AVG denotes the average MOTA excluding `uav0000268_05773_v`.

Tracker	IDF1(↑)	FAR(↓)	MT(↑)	ML(↓)	FP(↓)	FN(↓)	IDs(↓)	FM(↓)	MOTA(↑)	MOTP(↑)	Hz(↑)
V-IOU	56.1	0.76	297	514	11838	74027	**265**	1380	40.2	74.9	20
TrackCG [26]	**58.0**	0.86	323	395	14722	68060	779	3717	**42.6**	74.1	10
GOG_EOC [23]	46.5	**0.29**	205	589	**5445**	86399	354	**1090**	36.9	**75.8**	1
SCTrack [1]	45.1	0.39	211	550	7298	85623	798	2042	35.8	75.6	2.9
Ctrack [27]	51.9	1.95	**369**	**375**	36930	**62819**	1376	2190	30.8	73.5	15
FRMOT [24]	50.8	1.15	254	463	21736	74953	1043	2534	33.1	73.0	5
GOG [23]	45.1	0.54	244	496	10179	78724	1114	2012	38.4	75.1	**564.8**
IHTLS [9]	43.0	0.94	245	446	14564	75361	1435	2662	36.5	74.8	16.3
TBD [10]	45.9	1.17	302	419	22086	70083	1834	2307	35.6	74.1	0.7
H^2T [29]	44.4	0.95	214	494	17889	79801	1269	2035	32.2	73.3	1.56
CMOT [4]	51.3	1.42	282	435	26851	72382	789	2257	31.5	73.3	1.39
CEM [22]	19.2	1.12	105	752	21180	116363	1002	1858	5.1	72.3	7.74

Table 4. Multi-object tracking results on the VisDrone-VDT2018 testing set compared to state-of-the-art [31].

[6] and done for the experiments on UA-DETRAC, we perform an optimal assignment using the Hungarian algorithm as proposed in [5]. Although the computational complexity is increased, slightly better overall results are achieved for the VisDrone dataset.

For the validation set the MOTA improves by about an additional 1-2% whereas the number of ID switches is reduced by up to 60%. Accordingly, the ID based metrics IDF1, IDP and IDR improve consistently when using visual tracking. The Medianflow configuration has to be found to be the best general performing and is chosen for evaluation on the test set. Table 4 shows the final results in comparison to the state-of-the-art. A more detailed analysis can be found in [31] which reports V-IOU as the best tracker in terms of the average rank over all 10 metrics.

4. Conclusions

We presented an intuitive way of incorporating visual tracking into the intersection-over-union (IOU) tracker. In several experiments we showed that false negative detections can be compensated for robustly. As a result, the number of ID switches and fragmentations is reduced and thereby the quality of the tracks improves significantly. In the proposed tracking framework any visual single-object tracker can be used. When employing the popular KCF [14], frame-rates of over 200 fps can be achieved for high-definition video footage while outperforming the state-of-the-art on the UA-DETRAC and VisDrone datasets. This high-speed, high accuracy and simplicity of the proposed tracking approach makes it suitable for many use cases as well as a powerful baseline for future approaches.

5. Acknowledgements

The research leading to these results has received funding from BMBF-VIP+ under grant agreement number 03VP01940 (SiGroViD).

References

[1] N. M. Al-Shakarji, G. Seetharaman, F. Bunyak, and K. Palaniappan. Robust multi-object tracking with semantic color correlation. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–7, 2017.

[2] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1272, 2011.

[3] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1926–1933, 2012.

[4] S.-H. Bae and K.-J. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1218–1225, 2014.

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

[6] E. Bochinski, V. Eiselein, and T. Sikora. High-speed tracking-by-detection without using image information. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–6, 2017.

[7] Z. Cai, M. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In

IEEE International Conference on Computer Vision, pages 3361–3369, 2015.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition.*, pages 248–255, 2009.

[9] C. Dicle, O. I. Camps, and M. Sznaier. The way they move: Tracking multiple targets with similar appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2304–2311, 2013.

[10] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3d traffic scene understanding from movable platforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):1012–1025, 2014.

[11] R. Girshick. Fast r-cnn. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2015.

[15] Z. Kalal, K. Mikolajczyk, and J. Matas. Forward-backward error: Automatic detection of tracking failures. In *International conference on Pattern Recognition*, pages 2756–2759, 2010.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] T. Kutschbach, E. Bochinski, V. Eiselein, and T. Sikora. Sequential sensor fusion combining probability hypothesis density and kernelized correlation filters for multi-object tracking in video data. In *International Workshop on Traffic and Street Surveillance for Safety and Security at IEEE AVSS 2017*, pages 1–5, 2017.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37, 2016.

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[21] S. Lyu, M.-C. Chang, D. Du, L. Wen, H. Qi, Y. Li, Y. Wei, L. Ke, T. Hu, M. Del Coco, et al. Ua-detrac 2017: Report of avss2017 & iwt4s challenge on advanced traffic monitoring. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–7, 2017.

[22] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):58–72, 2014.

[23] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1201–1208, 2011.

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):1137–1149, 2017.

[25] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple people tracking by lifted multicut and person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3701–3710, 2017.

[26] W. Tian and M. Lauer. Fast cyclist detection by cascaded detector and geometric constraint. In *International Conference on Intelligent Transportation Systems*, pages 1286–1291, 2015.

[27] W. Tian and M. Lauer. Joint tracking with event grouping and temporal constraints. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–5, 2017.

[28] L. Wen, D. Du, Z. Cai, Z. Lei, M. Chang, H. Qi, J. Lim, M. Yang, and S. Lyu. UA-DETRAC: A new benchmark and protocol for multi-object tracking. *CoRR*, abs/1511.04136, 2015.

[29] L. Wen, W. Li, J. Yan, Z. Lei, D. Yi, and S. Z. Li. Multiple target tracking based on undirected hierarchical relation hypergraph. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1282–1289, 2014.

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

[31] P. Zhu, L. Wen, D. Du, X. Bian, H. Ling, Q. Hu, H. Cheng, C. Liu, X. Liu, W. Ma, Q. Nie, H. Wu, L. Wang, A. Schumann, D. Wang, D. Ortego, E. Luna, E. Michail, E. Bochinski, F. Ni, F. Bunyak, G. Zhang, G. Seetharaman, G. Li, H. Yu, I. Kompatsiaris, J. Zhao, J. Gao, J. Martinez, J. Miguel, K. Palaniappan, K. Avgerinakis, L. Sommer, M. Lauer, M. Liu, N. Al-Shakarji, O. Acatay, P. Giannakeris, Q. Zhao, Q. Ma, Q. Huang, S. Vrochidis, T. Sikora, T. Senst, W. Song, W. Tian, W. Zhang, Y. Zhao, Y. Bai, Y. Wu, Y. Wang, Y. Li, Z. Pi, and Z. Ma. VisDrone-VDT2018: The vision meets drone video detection and tracking challenge results. In *European Conference on Computer Vision*, pages 1–24, 2018.