# Self-supervised Polyp Re-identification in Colonoscopy

Yotam Intrator, Natalie Aizenberg, Amir Livne(✉), Ehud Rivlin, and Roman Goldenberg

Verily AI, Haifa, Israel
amirlivne@verily.com

**Abstract.** Computer-aided polyp detection (CADe) is becoming a standard, integral part of any modern colonoscopy system. A typical colonoscopy CADe detects a polyp in a single frame and does not track it through the video sequence. Yet, many downstream tasks including polyp characterization (CADx), quality metrics, automatic reporting, require aggregating polyp data from multiple frames. In this work we propose a robust long term polyp tracking method based on re-identification by visual appearance. Our solution uses an attention-based self-supervised ML model, specifically designed to leverage the temporal nature of video input. We quantitatively evaluate method's performance and demonstrate its value for the CADx task.

**Keywords:** Colonoscopy · Re-Identification · Optical Biopsy · Attention · Self Supervised

## 1 Introduction

Optical colonoscopy is the standard of care screening procedure for the prevention and early detection of colorectal cancer (CRC). The primary goal of a screening colonoscopy is polyp detection and preventive removal. It is well known that many polyps go unnoticed during colonoscopy [22]. To deal with this problem, computer-aided polyp detector (CADe) was introduced [13–16] and recently became commercially available [3]. The success of polyp detector sparkled the development of new CAD tools for colonoscopy, including polyp characterization (CADx, or optical biopsy), extraction of various quality metrics, and automatic reporting. Many of those new CAD applications require aggregation of all available data on a polyp into a single unified entity. For example, one would expect higher accuracy for CADx when it analyzes all frames where a polyp is observed. Clustering polyp detections into polyp entities is a prerequisite for computing such quality metrics as Polyp Detection Rate (PDR) and Polyps Per Colonoscopy (PPC), and for listing detected polyps in a report.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43904-9_57.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 590–600, 2023.
https://doi.org/10.1007/978-3-031-43904-9_57

One may notice that the described task generally falls into the category of the well known multiple object tracking (MOT) problem [26,27]. While this is true, there are a few factors specific to the colonoscopy setup: (a) Due to abrupt endoscope camera movements, targets (polyps) often go out of the field of view, (b) Because of heavy imaging conditions (liquids, debris, low illumination) and non-rigid nature of the colon, targets may change their appearance significantly, (c) Many targets (polyps) are quite similar in appearance. Those factors limit the scope and accuracy of existing frame-by-frame spatio-temporal tracking methods, which typically yield an over-fragmented result. That is, the track is often lost, resulting in relatively short tracklets (temporal sequences of same target detections in multiple near-consecutive frames), see Supplementary Fig.1.

A recently published method [2] addresses this limitation by combining spatial target proximity and visual similarity to match a polyp detected in the current frame to "active" polyp tracklets dynamically maintained by the system. The tracklets are built incrementally, by adding a single frame detection to the matched tracklet, one-by-one. However, this approach limits itself to use of close-in-time consistent detections, and cannot handle the frequent cases where polyp gets out of the field of view and long range association is required.

In this work we propose an alternative approach that allows polyp detections grouping over an extended period of time (up to 10 min), relaxing the spatio-temporal proximity limitation. It involves two steps: (I) a short-term multi-object tracking, which forms initial, relatively short tracklets, followed by (II) a longer-term tracklets grouping by appearance-based polyp re-identification (ReID). As the first step can be done by any generic multiple object tracking algorithm (e.g. we use a tracking by detection method [27]), in this paper we focus on the second step.

To avoid manual data annotation, which is extremely ineffective in our case, we turn to self-supervision and adapt the widely used contrastive learning approach [5] to video input and object tracking scenario.

As tracklet re-identification is a sequence-to-sequence matching problem, the standard solution is comparing sequences element-wise and then aggregating the per-element comparisons, e.g. by averaging or max/min pooling [21] - the so-called late fusion technique. We, on the other hand, follow an early fusion approach by building a joint representation for the whole sequence. We use an advanced transformer network [23] to leverage the attention paradigm for non-uniform weighing and "knowledge exchange" between tracklet frames.

We extensively test the proposed method on hundreds of colonoscopy videos and evaluate the contribution of method components using an ablation study. Finally, we demonstrate the effectiveness of the proposed ReID method for improving the accuracy of polyp characterization (CADx).

To summarize, the three main contributions of the paper are:

- An adaptation of contrastive learning to video input for the purpose of appearance based object tracking.
- An early fusion, joint multi-view object representation for ReID, based on transformer networks.
- The application of polyp ReID to boost the polyp CADx performance.

## 2 Methods

This work assumes the availability of an automatic polyp detector. Quite a few highly accurate polyp detectors were recently reported [14–16], detecting (multiple) polyps in a single frame. Our ultimate goal is to group those detections into sets corresponding to distinct polyps.

As briefly mentioned above, the proposed approach starts with an initial grouping of polyp detections using an off-the-shelf multiple object tracking algorithm. Such a tracker is expected to track polyps through consecutive frames as long as they do not leave the camera field of view, forming disjoint, time separated polyp tracklets. In this work we use the ByteTrack [27] "tracking by detection" algorithm, but, in principle, any other tracker could be used instead.

The resulting tracklets are typically relatively short, and there are quite a few tracklets corresponding to the same polyp. To improve the result, we propose an Appearance-based Polyp Re-Identification (ReID), which groups multiple disjoint tracklets by their visual appearance into a joint tracklet, associated with a single polyp. In what follows we describe in detail the proposed ReID component.

As stated above, the objective of ReID is to ascertain whether two time-separated, disjoint tracklets belong to the same polyp. To this end we seek a tracklet representation that allows measuring visual similarity between tracklets. The two basic alternatives are either a single representation for the whole tracklet, or a sequence of single-frame representations for each tracklet frame. We will consider both options below.

### 2.1 Single-Frame Representation for ReID

To generate a single frame representation we train an embedding model that maps a polyp image into a latent space, s.t. the vectors of different views of the same polyp are placed closer, and of different polyps away from each other [11].

A straightforward approach to train such model is supervised learning, which requires forming a large collection of polyp image pairs, manually labeled as same/not same polyp [1]. Such annotation turned out to be inaccurate and expensive. In addition, finding hard negative pairs is especially challenging, as images of two randomly sampled polyps are usually very dissimilar. Moreover, self-supervised techniques using extensive unannotated datasets has exhibited substantial advantages within the medical domain [12].

Hence, we turn to SimCLR [5], a contrastive self-supervised learning technique, which requires no manual labeling. In SimCLR the loss is calculated over the whole batch where all input samples serve as negatives of each other and positive samples are generated via image augmentations. Combined with the temperature mechanism this allows for hard negative mining by prioritizing hard-to-distinguish pairs, resulting in a more effective loss weighting scheme.

One caveat of SimCLR is the difficulty to generate augmentations beneficial for the learning process [5]. Specifically for colonoscopy, the standard image augmentations do not capture the diversity of polyp appearances in different views (see Fig. 1(c)).

Instead of customizing the augmentations to fit the colonoscopy setup, we leverage the temporal nature of videos, and take different polyp views from the same tracklet as positive samples (see Fig. 1(b)).

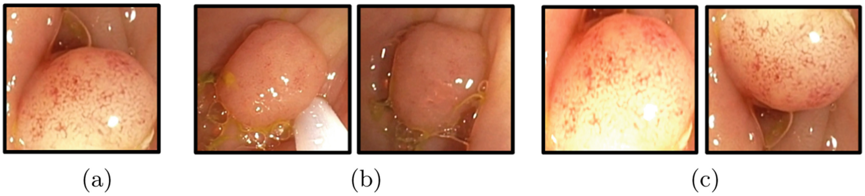

**Fig. 1.** (a) A polyp image, (b) two additional views of the polyp in (a) taken from the same tracklet, (c) two typical augmentations of the polyp in (a). Images in (b) offer more realistic variations, such as different texture, tools, etc.

Formally, a batch is formed by sampling one tracklet from $N$ different procedures to ensure the tracklets belong to different polyps. Two polyp views $i, j$ are sampled from each tracklet as positive pairs (same polyp). Let $f$ be the embedding model. The loss function for the positive pair $(i, j)$ is defined as:

$$\ell_{i,j} = -log \frac{exp(sim(f(i), f(j))/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} exp(sim(f(i), f(k))/\tau)} \tag{1}$$

where $sim$ is the dot product and $\tau$ is the temperature parameter [24]. The final loss is computed across all positive pairs in the batch.

Tracklets represented as sequences of per-frame embeddings can be matched by computing pair-wise distances between frames, followed by an aggregation - e.g. min/max/mean distance [4,10]. An example of similarities between frames can be seen in Supplementary Fig. 2.

### 2.2 Multi-view Tracklet Representation for ReID

As discussed earlier, an alternative to the single frame approach, is a unified representation for the whole tracklet. A commonly used practice is to compute single frame embedding (for each view) and fuse them [8,21], e.g. by averaging. The downside of those simple techniques is that they treat every frame in the same way, including bad quality, repeating, non-informative views. We postulate that learning a joint embedding of multiple views in an end-to-end manner will produce a better representation of the visual properties of a polyp, by allowing "knowledge exchange" between the tracklet frames.

To achieve this, we employ a transformer network [23], with the addition of BERT [7] classification token (CLS). The attention mechanism enables both frame based intra attention and selective weighting of the frames thus providing a more comprehensive tracklet representation. The overview of the architecture is presented in Fig. 2. Training this multi-view encoder is done similarly to training a single-view encoder using SimCLR, but now, instead of pairs of frames, we deal with pairs of tracklet.

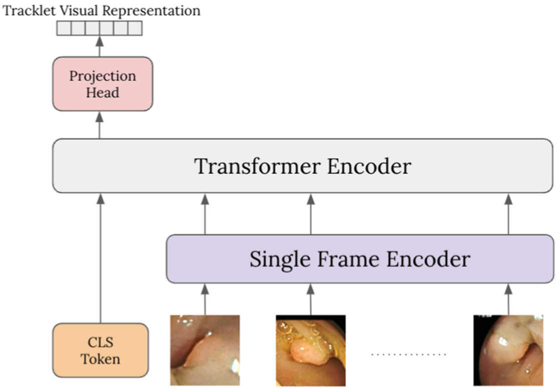

**Fig. 2.** Multi-view transformer encoder. Tracklet frames are passed through a single frame encoder to generate frame embedding. The embeddings then go through the transformer encoder, concatenated with the CLS token. Finally, the contextualized CLS token from the transformer encoder output goes through a projection head, resulting with the tracklet visual representation.

To generate positive tracklet pairs, we cannot apply the trick used for single frames, where positive pairs are sampled within the same tracklet. Instead we generate "pseudo positive" pairs from existing tracklets. We artificially split a tracklet into 3 disjoint segments, where the middle segment is discarded, and the first and the last segments are used as a positive pair, thus providing sufficiently different appearances of the same polyp as would happen in real procedures. In addition, this type of sampling approach, which effectively discards highly correlated samples from training, has been shown to improve model performance in [17].

## 3 Experiments

This section includes two parts. The first provides a stand-alone evaluation of the proposed ReID method. The second assesses the impact of ReID on polyp classification accuracy.

### 3.1 ReID Standalone Evaluation

**Dataset.** We use 22,283 colonoscopy videos, split into training (21,737) and test (546) sets. These recordings were captured from standard colonoscopy procedures conducted at six medical centers during the period of 2019 to 2022. The average length of the recorded procedures is 15 min, with a median duration of

13 min. For training, we automatically generated polyp tracklets using automatic polyp detection and tracking as described in Sect. 2.

The tracking algorithm might produce short and uninformative tracklets as well as outliers. The following clean up steps were performed on the training set: we filtered out tracklets shorter than 1 s or having less than 15 high confidence detections, as defined in [27], and took only the longest tracklet from every procedure. The thresholds were determined using analysis of the training set tracklets distribution. This yielded the training set of 15,465 tracklets (mean duration of 377 frames or 29 s). For evaluation, the test set polyp tracklets were manually annotated (timestamps and bounding boxes) by certified physicians. In addition, tracklet pairs from the same procedure were manually labeled as either belonging to the same polyp or not. This yielded 348 negative and 252 positive tracklet pairs.

**Training.** We utilize ResNet50V2 [9] as the single frame encoder, with an MLP head projecting the representation into a 128-dimensional embedding vector. We initialize the model using pre-trained ImageNet [6] weights. While ImageNet weights are not optimal for medical tasks [18,19], they offer training speedups [18]. The multi-view encoder consists of 3 transformer encoder blocks with an MLP projection head. We use LARS optimizer [25] with the learning rate of 0.01 and $\tau = 0.1$ as suggested in [5]. The batch size is set to 1024 for training both the single frame and the multi-view encoder.

We first train the single frame encoder and use its weights to initialize the single frame module of the multi-view encoder. Due to memory limitations, we use 8 views per tracklet during training, resulting in $1024 * 8 = 8192$ images per training step. The model was trained for 5,000 steps using cloud v3 TPUs with 16 cores. The single frame encoder has $24M$ parameters, and the multi-view encoder adds an additional $1M$ parameters.

**Evaluation.** We start by comparing various ReID techniques described in Sect. 2. Namely, we evaluate the accuracy of tracklet re-identification using: (a) single-frame representation with pairwise distances aggregation by Min / Max / Mean functions [4,10]; (b) multi-view representation by frame embeddings averaging; and, finally, (c) the joint embedding multi-view model. We evaluate the performance using AUC of the ROC and precision-recall curve (PRC) for tracklet similarity scores over the test set (see Table 1 and Supplementary Fig. 3). One can see that the joint embedding multi-view model outperforms all other techniques both on ROC and PRC.

In addition, we evaluate the effectiveness of ReID by measuring the average polyp fragmentation rate (FR), defined as the average number of tracklets polyps are split into. Obviously, lower fragmentation rate means better result (with the best fragmentation of 1), but it may come at the expense of wrong tracklet matching (false positive). We measure the fragmentation rate at the operating point of 5% false positive rate. The number of polyp fragments is determined by matching tracklets to manually annotated polyps and counting

**Table 1.** Polyp ReID accuracy for various ReID techniques.

| | Single-frame | | | Multi-view | |
|---|---|---|---|---|---|
| | Min | Max | Mean | Averaging | Joint Embedding |
| AUROC | 0.60 | 0.74 | 0.72 | 0.75 | 0.77 |
| AUPRC | 0.50 | 0.65 | 0.62 | 0.67 | 0.69 |

their number. Results presented in Table 2 demonstrate that ReID can reduce the fragmentation rate by over 50%, compared to a tracking only solution [27].

**Table 2.** Fragmentation Rate (FR) statistics before and after the ReID. FR STD is the FR standard deviation. Fragmented Polyps Ratio is the percentage of polyps divided into more than one tracklet.

| | FR | FR STD | Fragmented Polyps Ratio |
|---|---|---|---|
| Tracking | 3.3 | 3.3 | 0.64 |
| Tracking+ReID | 1.86 | 1.49 | 0.45 |

### 3.2 ReID for CADx

In this section, we investigate the potential benefits of using polyp ReID as part of a CADx system. Polyp CADx aims to assist physicians to figure out, in real time, during the procedure, whether the detected polyp is an adenoma.

Most reported CADx systems compute a classification score for each frame, and aggregate scores from multiple frames to determine the final polyp classification. Grouping polyp frames into a tracklet, to be fed into the CADx, is usually done by a spatio-temporal tracker [2]. Longer tracklets provide more information for polyp classification.

Here, we investigate if the proposed ReID model, used to group disjoint tracklets of the same polyp, can increase the accuracy of CADx.

**Data.** We use 3290 colonoscopy videos split into train, validation, and test sets (2666, 296, and 328 videos respectively). The videos are processed by a polyp detector and tracker to form polyp tracklets. The tracklets are then manually grouped together to build a single sequence for every polyp. Each polyp is annotated by a certified gastroenterologist as either adenoma or non-adenoma.

**CADx.** We trained a simple image classification CNN, composed of a MobileNet [20] backbone, followed by an MLP layer with a sigmoid activation, to predict the non-adenoma/adenoma score in $[0, 1]$, for each frame. The chosen architecture has 2.4M parameters and can run in real-time. The model was trained on Nvidia Tesla V100 GPU for 200 epochs with a learning rate of 0.001, using Adam optimizer.

For evaluation, we used the model to predict the classification score for each frame and aggregated the scores using soft voting to achieve the final prediction for each tracklet.

**Evaluation.** To assess the contribution of the ReID to polyp classification, we compare the CADx results on the test set, while using different grouping methods to merge multiple polyp detections into tracklets. The 3 evaluated methods are: (1) manual annotation (2) grouping by tracking, and (3) grouping by ReID. The manually annotated tracklets - the ground truth (GT) - are the longest sequences, containing all frames of each polyp in the test set. In grouping by tracking, we use tracklets generated by the spatio-temporal tracking algorithm [27]. Finally, for ReID, we merge disjoint tracklets by their appearance using the ReID model. By construction, tracklets generated by methods (2) and (3) are subsets of the corresponding manually annotated GT tracklet, and are assigned its polyp classification label. A visualization of the resulting tracklets using different grouping methods is provided in Supplementary Fig. 4. The number of resulting tracklets in the test set for each grouping method and polyp labels distribution are summarized in Table 3.

**Table 3.** CADx test data distribution and fragmentation rate (FR).

| Grouping | Tracklets | FR | Adenoma | Adenoma FR | Non-Adenoma | Non-Adenoma FR |
|---|---|---|---|---|---|---|
| Annotation | 608 | 1.0 | 464 | 1.0 | 144 | 1.0 |
| Tracking | 3161 | 5.20 | 2537 | 5.47 | 624 | 4.33 |
| Tracking+ReID | 1023 | 1.68 | 813 | 1.75 | 210 | 1.46 |

We ran the CADx model on tracklets generated by the 3 grouping methods. We compute the $F_1$ score and the AUC for the tracklet classification task. In addition, we measure the CADx sensitivity at specificity=0.9. The results are summarized in Table 4. The result on the manually annotated data is the accuracy upper-bound and is brought as a reference point. One can see that the ReID based approach significantly improves the CADx accuracy compared to the tracking-based grouping.

**Table 4.** Optical biopsy result per grouping method.

| Grouping | AUC | F1 (Macro) | F1 (Micro) | Sensitivity @ Specificity=0.9 |
|---|---|---|---|---|
| Annotation | 0.95 | 0.88 | 0.91 | 0.86 |
| Tracking | 0.86 | 0.77 | 0.83 | 0.71 |
| Tracking+ReID | 0.90 | 0.82 | 0.88 | 0.79 |

## 4 Conclusions

In this study we present a novel multi-view self-supervised learning method for learning informative representations of a sequence of video frames. By jointly encoding multiple views of the same object, we get more discriminative features in comparison to traditional embedding fusion techniques. This approach can be used to group disjoint tracklets generated by a spatio-temporal tracking algorithm based on their appearance, by measuring the similarity between tracklets representations. Its applicability to medical contexts is of particular relevance, as medical data annotation often requires specific expertise and may be costly and time consuming. We use this method to train a polyp re-identification model (ReID) from large unlabeled data, and show that using the ReID model as part of a CADx system enhances the performance of polyp classification. There are some limitations however in identifying polyps based on their appearance, as it may be changed drastically during the procedure (for example, during resection). In future work we may examine the use of ReID for additional medical applications, such as listing detected polyps in an automatic report, bookmarking of specific areas of the colon during the procedure, and calculation of clinical metrics such as Polyp Detection Rate and Polyps Per Colonoscopy.

## References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015)
2. Biffi, C., Salvagnini, P., Dinh, N.N., Hassan, C., Sharma, P., Cherubini, A.: A novel ai device for real-time optical characterization of colorectal polyps. NPJ Digital Med. **5**(1), 84 (2022)
3. Brand, M., et al.: Frame-by-frame analysis of a commercially available artificial intelligence polyp detection system in full-length colonoscopies. Digestion **103**(5), 378–385 (2022)
4. Breckon, T.P., Alsehaim, A.: Not 3d re-id: simple single stream 2d convolution for robust video re-identification. In: 2020 25th International conference on pattern recognition (ICPR), pp. 5190–5197. IEEE (2021)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)

6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Gao, J., Nevatia, R.: Revisiting temporal modeling for video-based person reid. arXiv preprint arXiv:1805.02104 (2018)
9. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
10. He, T., Jin, X., Shen, X., Huang, J., Chen, Z., Hua, X.S.: Dense interaction learning for video-based person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1490–1501 (2021)
11. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
12. Hirsch, R., et al.: Self-supervised learning for endoscopic video analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2023)
13. Lachter, J., et al.: Novel artificial intelligence-enabled deep learning system to enhance adenoma detection: a prospective randomized controlled study. iGIE (2023)
14. Livovsky, D.M., et al.: Detection of elusive polyps using a large-scale artificial intelligence system (with videos). Gastrointest. Endosc. **94**(6), 1099–1109 (2021)
15. Ou, S., Gao, Y., Zhang, Z., Shi, C.: Polyp-yolov5-tiny: a lightweight model for real-time polyp detection. In: 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), vol. 2, pp. 1106–1111. IEEE (2021)
16. Pacal, I., Karaboga, D.: A robust real-time deep learning based automatic polyp detection system. Comput. Biol. Med. **134**, 104519 (2021)
17. Qian, R., et al.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6964–6974 (2021)
18. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: understanding transfer learning for medical imaging (2019)
19. Rajpurkar, P., et al.: Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning (2017)
20. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv 2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
21. Seeland, M., Mäder, P.: Multi-view classification with convolutional neural networks. PLoS ONE **16**(1), e0245230 (2021)
22. Van Rijn, J.C., Reitsma, J.B., Stoker, J., Bossuyt, P.M., Van Deventer, S.J., Dekker, E.: Polyp miss rate determined by tandem colonoscopy: a systematic review. Official J. Am. College Gastroenterology| ACG **101**(2), 343–350 (2006)
23. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)
24. Wang, F., Liu, H.: Understanding the behaviour of contrastive loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2495–2504 (2021)

25. You, Y., et al.: Large batch optimization for deep learning: training bert in 76 minutes. arXiv preprint arXiv:1904.00962 (2019)
26. Yu, T., et al.: An end-to-end tracking method for polyp detectors in colonoscopy videos. Artif. Intell. Med. **131**, 102363 (2022)
27. Zhang, Y., et al.: Bytetrack: multi-object tracking by associating every detection box. In: Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022, Proceedings, Part XXII, pp. 1–21. Springer (2022). https://doi.org/10.1007/978-3-031-20047-2_1