Turkish Journal of Electrical Engineering and Computer Sciences

Volume 31 | Number 2 Article 4

1-1-2023

Improved Object Re-Identification via More Efficient Embeddings

ERTUGRUL BAYRAKTAR

Follow this and additional works at: https://journals.tubitak.gov.tr/elektrik

 Part of the Computer Engineering Commons, Computer Sciences Commons, and the Electrical and Computer Engineering Commons

Recommended Citation

BAYRAKTAR, ERTUGRUL (2023) "Improved Object Re-Identification via More Efficient Embeddings," *Turkish Journal of Electrical Engineering and Computer Sciences*: Vol. 31: No. 2, Article 4. https://doi.org/10.55730/1300-0632.3984

Available at: https://journals.tubitak.gov.tr/elektrik/vol31/iss2/4

This Article is brought to you for free and open access by TÜBİTAK Academic Journals. It has been accepted for inclusion in Turkish Journal of Electrical Engineering and Computer Sciences by an authorized editor of TÜBİTAK Academic Journals. For more information, please contact academic.publications@tubitak.gov.tr.

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2023) 31: 282 – 294
© TÜBİTAK
doi:10.55730/1300-0632.3984

Improved object reidentification via more efficient embeddings

Ertugrul BAYRAKTAR*
Department of Mechatronics Engineering, Yıldız Technical University, Beşiktaş, İstanbul, Turkey

Received: 29.06.2022 • **Accepted/Published Online:** 21.01.2023 • **Final Version:** 23.03.2023

Abstract: Object reidentification (ReID) in cluttered rigid scenes is a challenging problem especially when same-looking objects coexist in the scene. ReID is accepted to be one of the most powerful tools for matching the correct identities to each individual object when issues such as occlusion, missed detections, multiple same-looking objects coexisting in the same scene, and disappearance of objects from the view and/or revisiting the same region arise. We propose a novel framework towards more efficient object ReID, improved object reidentification (IO-ReID), to perform object ReID in challenging scenes with real-time processing in mind. The proposed approach achieves distinctive and efficient object embedding via training with the triplet loss, with input from both the foreground/background split by bounding box, and the full input image. With extensive experiments on two datasets serving for Object ReID, we demonstrate that the proposed method, IO-ReID, obtains a higher ReID accuracy and runs faster compared to the state-of-the-art methods on object ReID.

Key words: Object reidentification, image retrieval, triplet loss, embedding generation, ranking

1. Introduction

Object reidentification (ReID) addresses the task of retrieving the image of an object of interest (i.e. a probe) from the images of a set of known objects (i.e. gallery) captured from multiple camera viewpoints. Given a probe person-of-interest, the goal of ReID is to determine whether this person has appeared in another place at a distinct time captured by a different camera [1]. The probe person can be represented by an image, a video sequence, and even a text description. Due to the urgent demand of public safety and increasing number of surveillance cameras in university campuses, theme parks, streets, etc., person ReID is imperative in intelligent video surveillance system designs. Given its research impact and practical importance, ReID is a fast-growing vision community. Object ReID in rigid scenes plays an essential role for vision tasks, such as consistent object tracking and object-based simultaneous localization and mapping (SLAM) [2, 3], where the capability of assigning the unique identity to each object detected from varying view points is desired. A large corpus of research can be found on human/person/pedestrian ReID, a closely related problem whose focus is on the retrieval of person from nonoverlapping camera views [4–6]. However, the study on object ReID still remains limited and much less explored [7, 8].

Object ReID with only visual cues of the detected object can be challenging because i) objects are occluded due to scene clutters and ii) similar-looking objects or multiple instances of the same category exist in the same environment. One of the earlier approaches on the issue, re-OBJ, makes use of the visual background (bg) cues by proposing combined deep features extracted from the foreground (fg) and background (bg), which are

*Correspondence: eb@yildiz.edu.tr

This work is licensed under a Creative Commons Attribution 4.0 International License.

segmented by a Mask R-CNN [9] model, to train a triplet network for the job of object ReID in static scenarios. The combined feature embedding proves to improve the the ReID accuracy compared to using features extracted from only fg. However, re-OBJ is not designed with the processing efficiency in mind, with a running speed less than 4 Hz on a standard PC, leading to undesired compromise on real-time performance if integrated into a complete object tracking system.

In this work, we investigate the efficiency of the major components of re-OBJ, and propose IO-ReID, a refined version of re-OBJ, by improving both segmentation backbone and the embedding generation module, achieving a higher ReID accuracy at a faster speed. More specifically, we propose three main modifications on top of re-OBJ: Firstly, we replace the format of segmentation from the mask to the bound boxes (BBXes), which achieves a significant improvement in ReID accuracy and also allow us to replace the most time-consuming module, Mask R-CNN, to other faster object detectors. Moreover, to improve the efficiency in embedding generation, we eliminate the most time-consuming element of re-OBJ, i.e. the identical convolutional blocks (ICBs), but at the cost of a ReID performance reduction. To compensate such performance drop, we further include the full image together with fg/bg splits to the composition of the visual features, achieving a slightly higher accuracy without compromising much the speed. With extensive experimental studies, we justify our proposed modifications, prove the advantages of IO-ReID in terms of accuracy and speed in comparison to re-OBJ and related methods in both accuracy and speed with datasets for object ReID task.

The rest of the paper is organized as follows: Section 2 provides the related work on object ReID. Section 3 explains IO-ReID with detailed description for each stage. Section 4 gives the ablation studies and baseline comparisons as a result of real scenes. Finally, Section 5 concludes the paper by providing perspectives for future work.

2. Motivation and related work

Since many object ReID algorithms start with distinguishing the object of interest from the rest of the content of the scene phase, we also give place to object detection and instance segmentation methods. Afterwards, we will cover object ReID and the closely-related image retrieval approaches. Recently, deep-learning-based object detection [10–13] and instance segmentation [14–16] is composed of feature extraction, region proposal, and prediction.

2.1. Object detection and instance segmentation

The key phases in object detection/segmentation are as follows: *i)* feature extraction using an existing model, *ii)* region proposal trough the features, and *iii)* prediction heads made up of fully connected (FC)-layers classifying the region proposals. In order to collect features necessary for their prediction heads, detection/segmentation models commonly employ models initially intended for object classification, such as AlexNet [17], VGGNet [18], inception [19], ResNet [20], or a model derived from one of these models. For instance, YoloV3 [21] employs a modified ResNet while earlier Yolo-based models used modified VGGNet and AlexNet, respectively. Mask R-CNN, an instance segmentation model, is built on top of faster R-CNN [22] where ResNet is the backbone model and it also has a mask segmentation part. As being a more recent instance segmentation model YOLACT++ [23] can run in real-time, where the speed performance is achieved by linearly combining the two outputs to produce a parallel prototype mask and mask coefficient predictions for each instance before generating an instance mask. However, the accuracy and robustness of Mask R-CNN appear to be significantly better than other techniques.

2.2. Image retrieval

The main goal for object ReID is to find the exact same match of objects from a set of images, i.e. gallery, for the given input image that contains the relevant content [7] whilst image retrieval aims to associate the most similar instance(s) from a large database of images with the input image [24–26]. The image retrieval approach introduced in [24] employs fixed-length features extracted from a particular region via a fine-tuned network for landmark images. GEM [25] and GEM(AP) [26] claim to increase performance through data variation by structure from motion and training by list-wise ranking loss. Similarly, [27] introduces a pipeline localizing targets as feature maps as in [28], besides [29] is an approach that provides basic spatial reranking as a result of feature proposal network for analyzing the effects of fine-tuning. Additionally, [30] claims improvement by using regional aggregations of deep features.

If a scene contains multiple instances of the same category that appear together, image retrieval most likely fails. The re-OBJ [7] introduces an object ReID pipeline to predict the exact same object with the probe. re-OBJ uses mask-RCNN to split detected instances from the bg and extracts features from these two sources separately by Siamese ResNet50 giving embedding. Then with these joint features, re-OBJ trains a triplet network yielding final outputs, which are ranked to give results. Our work, *IO-ReID*, is mostly related to [7], with the main efforts to improve its real-time performance. On the other hand, object ReID has also been incorporated into more complete tracking frameworks as in [31–34].

2.3. ReID

In particular, ReID algorithms can handle if multiple object instances amongst the same object category exist together in the same view unlike image retrieval methods. ReID is widely considered by means of a person/pedestrian issue for which the appearance-based features [35] play an important role. These features are obtained through *i)* extraction and refinement processes for improved object embeddings [36], and *ii)* comparison of queries and gallery images to provide metric-driven ranking [37]. There are also other approaches using local classical feature symmetries for comparison with the gallery images through Euclidean metric [35, 38]. On the contrary, in [39], the visual features of person as the fg object are enhanced via temporal information, which matches visual cues according to minimum distances between them. For more information on the most recent developments in human ReID, readers can refer to [6]. Following the obtaining of spatial information re-OBJ method achieves the final embedding for the triplet loss layer using the combined representation of both the bg and fg in order to train a triplet loss function. In addition, object ReID is further adopted into more complicated tasks like object tracking as in [32–34]. The approach introduced in [31] aims object Re-ID for object segmentation purpose by subtracting the sequential frames. Thanks to utilizing a learned matching function, the authors in [32] present a ReID-driven tracker that processes over 4096-dimension embedding in consequence to being trained on an extensive dataset and hence can be even used in unseen targets without modification. The studies in [33, 34] present object ReID techniques, which are driven by reranking, for object segmentation and tracking in video sequences.

3. Improved object ReID

The capability of assigning the unique identity to each object detected in a video stream plays an essential role in many vision-based tasks, such as object tracking or 3D object localization. Our proposed *IO-ReID* is composed of three modules as shown in Figure 1. Given an anchor (A) image (or probe image), we first apply YoloV3 as

the object detection module with the fg segmented in the format of BBX. We then pass the segmented fg/bg and the full image to ResNet50 to extract their corresponding visual features, which are then concatenated and passed to the pre-trained triplet network to produce a more discriminative embedding for the final inference.

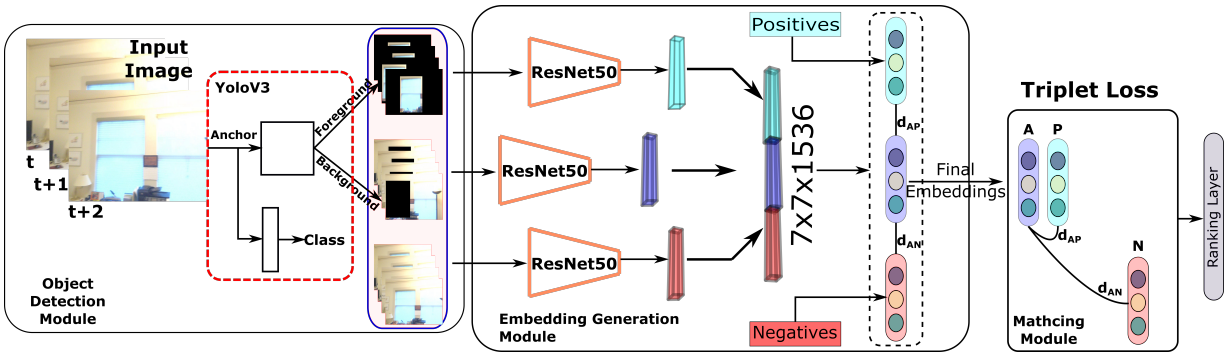


Figure 1. *IO-ReID* is composed of three components as instance segmentation, embedding generation, and matching. Anchors (A) are separated as bg/fg, which are next passed through parallel ResNet50s with the full images to generate embeddings the same as P and N. Final embeddings are used to train the triplet network of object matching module. Embedding generation and object matching processes require offline training, while the complete pipeline is used for online inference.

During training, the triplet network aims to generate an embedding space where each A is close to its Positive (P), the image with the same object captured from another view, while staying far from the Negative (N), the image without the same object. We provide more detailed description on the embedding generation section. During tests, once we split each detected instance as bg/fg, we extract features to get embeddings that are used for triplet comparison. In the last stage, reranking based on the Euclidean distance is applied to find the best match.

3.1. Embedding generation

The proposed embedding generation module exploits pretrained ResNet50 for visual feature extraction as in re-OBJ [7]. Moreover, we propose three main differences on top of the re-OBJ by properly adjusting elements considering their trade-off between the speed and accuracy, leading to a faster and more discriminative embedding. Firstly, we replace the time-consuming ICBs in re-OBJ for encoding spatial information with simple fully connected layer, which also enables us to adjust dimension. However, the removal of ICBs leads to a performance drop in the ReID accuracy. To compensate, we include an additional feature which is extracted from the full image into the embedding, which results in a more discriminative embedding with additional information capturing the relation between fg and bg. Herein, for each of the fg, bg, and full images, the embeddings are formed by concatenating the features coming from ResNet50s instead of ICBs.

Finally, we exploit the BBX as the segmentation format instead of the masks, which helps to obtain a more discriminative embedding. The choice of BBX also allows us to flexibly substitute the instance segmentation module with any faster object detectors.

The concatenated embeddings are employed to discriminate triplets as shown in Figure 2, where input, A, is aimed to be placed as close as possible to P and as far as possible from N. During tests, once we split each detected instance as bg/fg, we extract features to get embeddings that are used for triplet comparison. In the last stage, reranking based on the Euclidean distance is applied to find the best match.

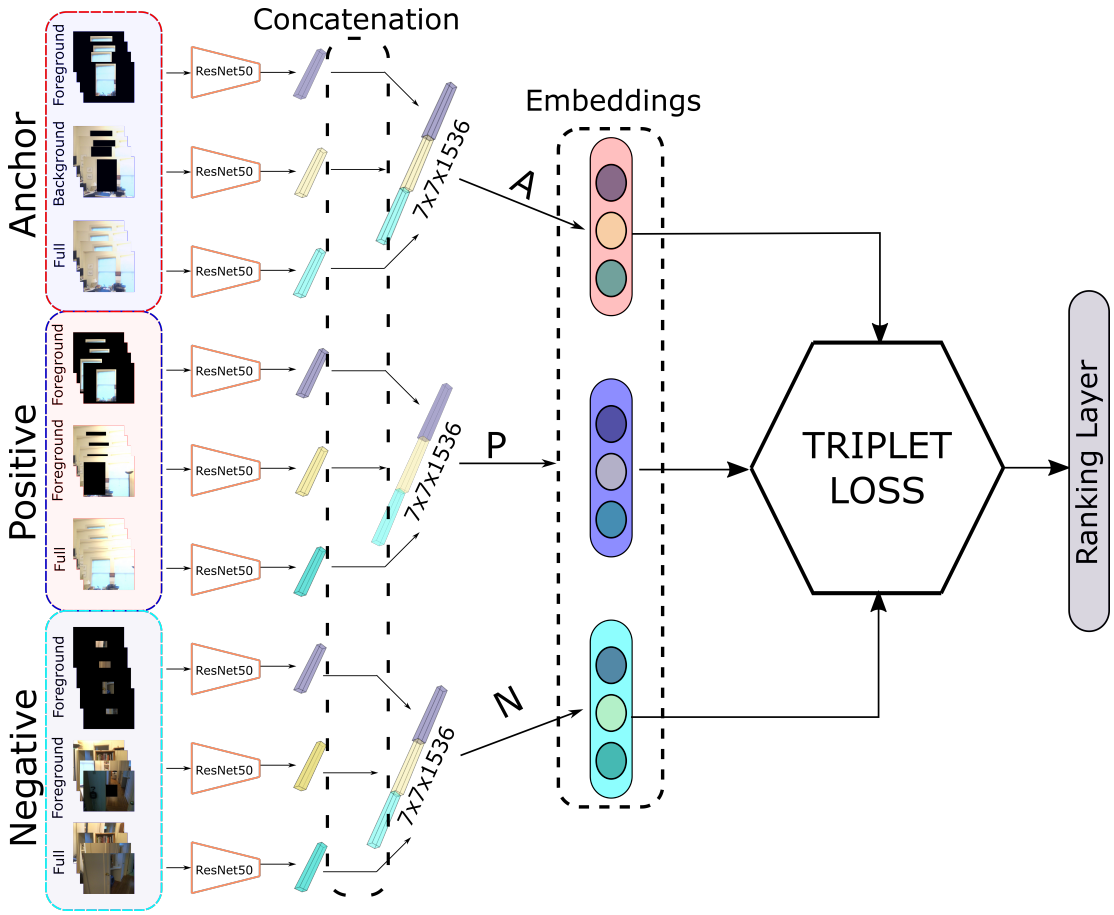


Figure 2. Embedding generation through concatenation of fg, bg, and full images for triplet-based matching.

3.2. Triplet-based object matching and ranking

With the concatenated features extracted by parallel ResNet50s from fg/bg and the full image, we generate the final embedding by training a triplet network. The objective of the training is to bring P of the triplets close to A and force N far from A. For a set of triplet images, I_{iA}, I_{iP}, and I_{iN}, we aim to achieve an embedding function that minimizes distance between A and P (vice versa for A-N pair) such that:

$$D(I_{iA}, I_{iP}) < D(I_{iA}, I_{iN}),$$

where D gives the squared Euclidean distance. We define triplet ranking loss function as follows:

$$\mathscr{L}(I_{iA}, I_{iP}, I_{iN}) = max\left[0, M + D(I_{iA}, I_{iP}) - D(I_{iA}, I_{iN})\right],$$

where M is the triplet margin adjusting the pairwise distances.

At inference time, we get the ranked probabilistic outputs correspondingly as a result of the triplet-based ranking loss at which every input is compared with all gallery using Euclidean distance. The outputs are ranked only if a correct match occurs.

To learn the similarity represented by $s_{i,j} = s(I_i, I_j)$, we seek for higher s scores for similar images.

4. Experiments

We evaluate IO-ReID with datasets containing object instances segmented with both ground-truth annotations and object detector outputs. We conduct experiments on both as ablation study to justify our design choices, and comparison to the state-of-the-art (sota) methods for addressing object ReID and the tightly related image retrieval task.

Performance metric: We follow the common convention for evaluation of ReID methods with mean average precision (mAP) and accuracy rates for rank-N matches by a cumulative matching characteristic (CMC) curve. We principally take the match into account only the exact instance as a correct match and the matches to similar objects are accepted as wrong matches.

Training details: We trained the model for 500 epochs with the learning rate scheduled to decrease exponentially. The margin for triplet loss is 0.2 and Adam [40] is used as the optimizer. The dimension of the feature extracted from one of the triplets (i.e. A, P, or N) through ResNet50 appears to be 7x7x512 whereas the joint feature becomes 7x7x1536 and the final embedding dimension is set to be 4096. All experiments are run on a standard machine with an i7-9750 CPU running at 2.60 GHz and a GPU of Nvidia GTX 1080ti on Ubuntu 18.04 LTS.

4.1. Data preparation

We used ReObj-MR from re-OBJ where Mask R-CNN is the instance segmentation backbone. In other words, the Mask R-CNN splits the full images into fg/bg pairs. ReObj-MR contains 707 instances selected from 27 scenes from ScanNet [41]. Furthermore, we employ ReObj-GT, containing 32 scenes with 849 instances, where the detections are extracted from 3D annotations provided in ScanNet. This information is exploited to get ground-truth BBXes for the selected objects. Once BBXes are obtained, then fg/bg pairs are split accordingly. It is vital that the IDs for each annotated object in ReObj-GT are scene-specific as in the original ScanNet dataset. For example, an object with an ID of "1" can represent a particular object with the category "chair" in the scene of interest (e.g., scene0054), but the ID "1" can represent a different object with the category "TV" in another scene (e.g., scene0061). By convention, images are discarded in ReOBJ-GT if any of the objects is obscured by another object by more than 15% or the BBXes are too small or do not belong to the secene of interest. We arranged our dataset as triplets for the proper learning of a pairwise ranking model as shown in Figure 3, where the BBXes of ReObj-GT and masks of ReObj-MR for the same table from the same view are given comparably. We set probes to be the individual objects whereas positives in the gallery to be the same with probes with different views and negatives to be different from both.

4.2. Ablation study

We perform experiments using different variants of our proposed pipeline regarding the structure and embedding choices to justify *IO-ReID*: *i)* **IO-ReID-D**: re-OBJ removes the ICBs; *ii)* **IO-ReID-DF**: **IO-ReID-D** adds the feature from the full image to the feature composition; *iii)* **IO-ReID-DFB** is **IO-ReID-DF** uses the BBX to segment fg/bg instead of masks; *iv) IO-ReID*: **IO-ReID-DFB** replaces the object detector from Mask-RCNN to YoloV3. The structure and embedding choices for the variants are summarized in Table 1 accordingly. For a fair comparison, we performed the training of the above-mentioned configurations separately on the two sets of data, i.e. ReOBJ-MR and ReOBJ-GT.

Table 2 shows the ReID performance and the decomposed time performance for each module, namely instance segmentation backbone (ISB), embedding generation (EG), and object matching (OM), respectively.

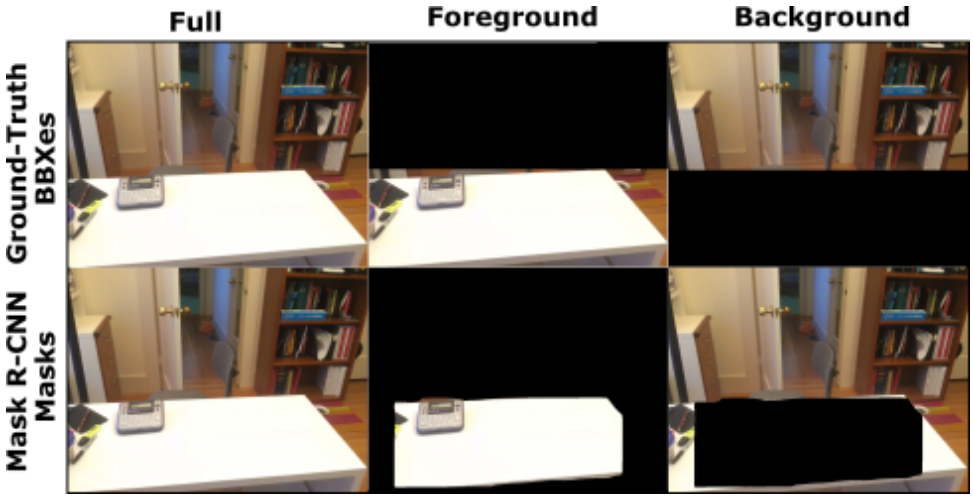


Figure 3. The example appearances of BBXes from ReObj-GT and masks from ReObj-MR for the same sample object for the same view.

The theory suggests [42, 43] that mAP values lie in the interval of [0, 1], which are projected to be in the interval of [0, 100], hence are considered percentages in this study. In addition, the threshold for mAP is used to specify the acceptable ratio of the intersection between the predicted outputs and ground-truth results, which is herein consistently set to be 0.25 to minimize the effects of object detection. By removing ICBs, IO-ReID-D makes the EG and OM faster, while the mAPs decreases significantly. By further including the full images together with the explicit split bg/fg images for EG, we see that IO-ReID-DF can remarkably enhance the performance, surpassing re-OBJ. By further replacing masks with BBXes, IO-ReID-DFB improves the ReID accuracy. Finally, when we replace mask R-CNN of ISB with YoloV3, *IO-ReID* achieves even a higher speed thanks to the faster ISB module, with a slight accuracy drop impacted by the detector performance, yet the mAP is still higher than re-OBJ.

Table 1. The structure and embedding choices for the variants of IO-ReID.

Method name	**Structure and embedding choice**
IO-ReID-D	Remove ICB from re-OBJ.
IO-ReID-DF	Add features from full images to the IO-ReID-D feature composition.
IO-ReID-DFB	Replace masks of IO-ReID-DF with BBXes to segment fg/bg.
IO-ReID	Replace object detector of IO-ReID-DFB with YoloV3.

Table 2. Results with various versions of *IO-ReID* evaluated on dataset ReObj-MR and ReObj-GT.

Method	**mAP [%] regarding datasets**		Time [fps]			
	ReObj-MR	ReObj-GT	ISB	EG	OM	Total
re-OBJ	67.8	**54.4**	**4.5**	19.4	41.3	3.36
IO-ReID-D	62.4	52.3	**4.5**	**72.4**	**107.6**	4.08
IO-ReID-DF	68.1	54.34	4.5	16.8	**107.6**	3.44
IO-ReID-DFB	**70.12**	**56.1**	4.5	16.8	**107.6**	3.44
IO-ReID	69.6	55.8	**10.7**	16.8	**107.6**	**6.16**

In addition, we do notice that the performance measures using ReOBJ-GT dataset are lower than ReOBJ-MR, which is because bg/fg splits performed by mask-RCNN are noisy and the segments contain wrongly labeled objects and partial masks, causing unreliable performance measurement. Instead, ReOBJ-GT contains correct labeling using the GT annotations from ScanNet.

4.3. Comparison with baseline methods

Herein, the performance of the proposed model, *IO-ReID*, is compared with re-OBJ, GEM, GEM(AP), and R-ASMK to justify its advantages and success on challenging ReID tasks. GEM, GEM(AP), and R-ASMK are originally instance search methods while re-OBJ is an object ReID algorithm. GEM(AP) is an extension of GEM, which improves performance through increasing data reconstruction. GEM(AP) further employs a list-wise ranking loss with a multistaged optimization configuration whilst training. Besides, R-ASMK first extracts local features and then combines them to selectively match filters within a two-staged framework. Following the examination of inner structures of retrieval methods, which consist of feature extraction and embedding generation as well as ranking modules, their suitability to ReID task becomes clear. Therefore, the aforementioned approaches are modified to work in a ReID setup appropriately herein. For a fair comparison, the feature extraction backbones of GEM, GEM (AP), and R-ASMK are further fine-tuned by using ReObj-GT. The ICB structure of re-OBJ is likewise trained by replacing ReObj-MR with ReObj-GT to get the results for ReObj-GT dataset.

Results obtained from running the compared methods on both datasets in terms of mAP and speed are given in Table 3. With our performance measure which only accounts for the match of exact object instance, although they are carefully adapted to ReID task for a fair comparison, we do notice that the image retrieval methods are worse than the frameworks that are originally developed for object ReID task. The basis beneath this poor performance is because the image retrieval algorithms are primarily structured for finding similar images from a set of database images. For this reason, the increase in the performance evaluation metric causes a decrease in the matching outputs. The instance retrieval methods appear to be slower than our framework because of their complicated architectures. Besides, it appears to be that the features generated by the backbones of all methods play an important role in making an approach more successful on ReID tasks.

Table 3. Results with *IO-ReID* and the compared sota methods evaluated on ReObj-MR and ReObj-GT.

Method	**mAP [%] regarding datasets**		Time
	ReObj-MR	ReObj-GT	[fps]
GEM [25]	54.7	50.8	1.7
GEM (AP) [26]	55.9	52.4	1.03
R-ASMK [30]	61.3	54.0	1.41
re-OBJ [7]	67.8	54.4	3.36
IO-ReID	69.6	55.8	**6.16**

As a result of comparing the estimations with anchor and calculating an ordered list considering the probability values, CMC curves are obtained as illustrated in Figure 4, which further give mAP over the ranks. Top-N accuracy, in other word examining if the top-N images within the ordered gallery contain an exact match with the anchor image, is taken into account. It is obvious that the performance outputs of re-OBJ and *IO-ReID* progress better on both datasets, namely ReOBJ-MR and ReOBJ-GT, where the increases also

occur remarkably sharper than that achieved by other methods. We show the behaviors for top-20 predictions because the curves for all compared methods start to converge horizontal axis remarkably after Rank-20 for both dataset, where the gap between the saturated accuracy of *IO-ReID* and other methods on ReOBJ-GT is larger than ReOBJ-MR.

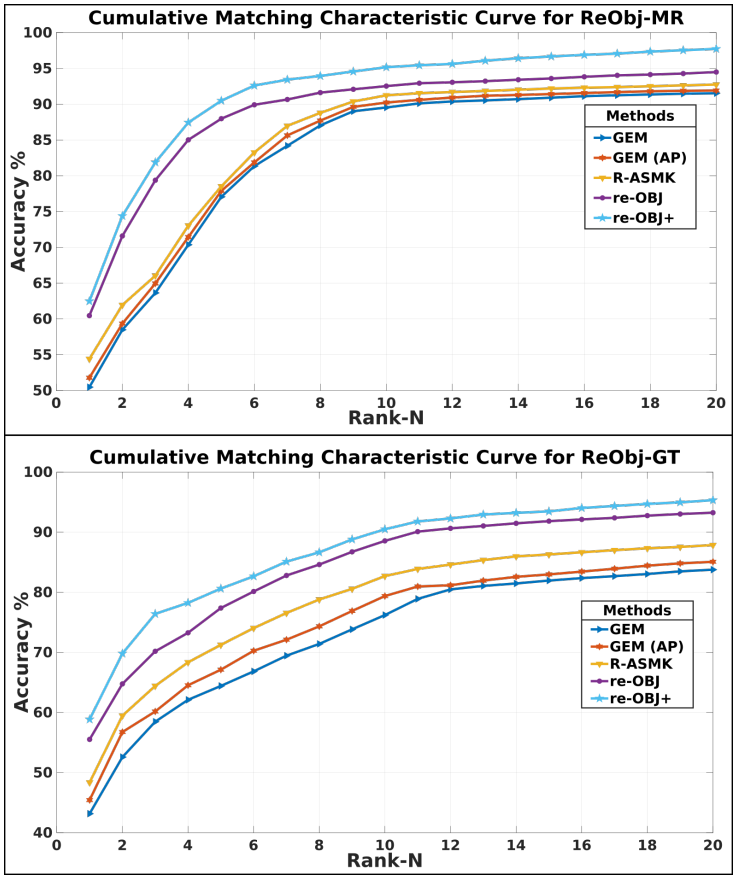


Figure 4. The comparison of progress of CMC curves for all the methods that are obtained due to experiments on ReOBJ-MR and ReOBJ-GT, respectively.

The main reason behind the fact that the compared methods start from higher accuracy levels and also end-up at greater values is that the samples of ReOBJ-MR dataset are noisy and have not been filtered out. For instance, if a wrong label is assigned to any object segmented out by mask R-CNN and the object detectors of the compared algorithms also assigned the same wrong label to that object, then these kinds of matches are assumed to be correct in the experiment phase, which result in increase in the performance. For this reason, testing the algorithms with the ReOBJ-GT is vital to have robust and reliable insights about their real performance. The results from both datasets nevertheless confirm each other and also show the importance of testing in multiple datasets through highlighting the differences.

Finally, Figure 5 shows several qualitative results for retrieving the exact instance within the top N ranks, with the *cup* as the first probe and *couch* as the second and third probes.

We apply *IO-ReID* for the first and second probes, whilst IO-ReID-DFB for the third probe. *IO-ReID* correctly predicts the exact same *cup* amongst similar instances; however, it fails to predict the correct *couch*. In the second test, the predicted couch with the highest score, which is shown in the most-left image of

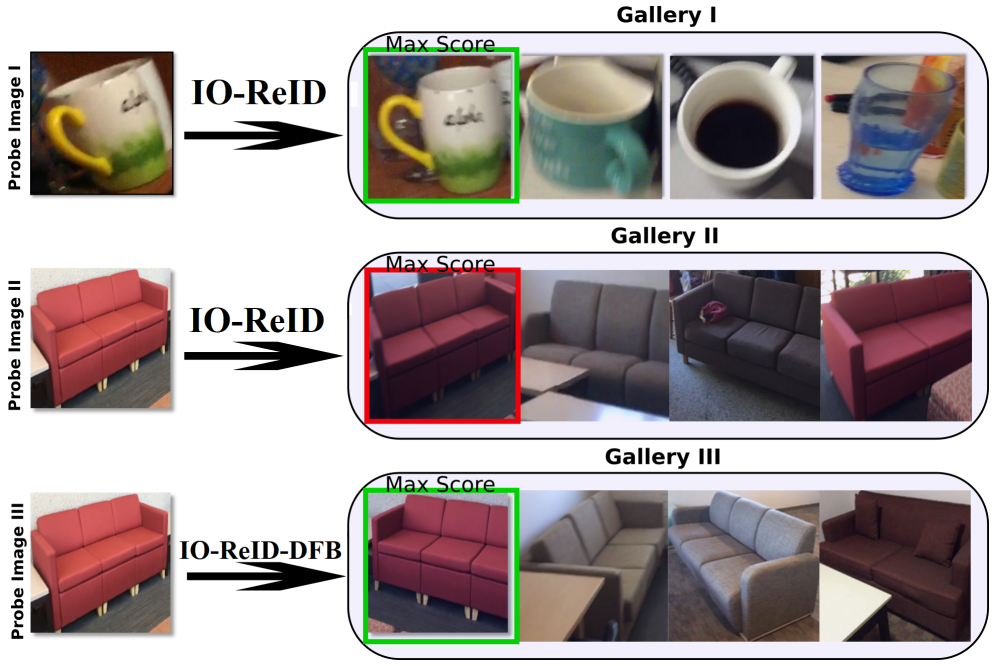


Figure 5. Qualitative results of Rank-N predictions for anchor object, *cup* as the first probe instance and *couch* as the second and third, amongst gallery images for which *IO-ReID* is utilized for the first and second rows and re-OBJ-DFB is applied for the third row.

the Gallery II, also contains a couch with similar appearance to the probe image, but the background exposes that the predicted couch takes part in a different surrounding the probe. On the other hand, IO-ReID-DFB predicts the exact same instance with a slower processing speed, despite the angle of view of predicted the couch is remarkably different in the third experiment.

5. Conclusion

We proposed an efficient framework to address the task of object ReID in rigid scenes. Through extensive experiments, we prove that our proposed framework, *IO-ReID*, can run in near real-time while maintaining the ReID accuracy in challenging scenes compared to the sota methods. Additionally, our analysis clearly unveils the unsuitability of image retrieval techniques in ReID tasks. There are, nevertheless, limitations for a typical object ReID approach due to missed detections, variations in illumination, occlusions, and cluttered background. *IO-ReID*is additionally influenced by same-looking objects that appear in the same scene and changes in the background. We managed to compensate the accuracy drop at IO-ReID-D by using bg, fg, and full images together to depict a more complete relationship with the target object and its neighboring pixels. As the structure and embedding choices for the *IO-ReID*variants reveal the relations between each approach, the detailed results clearly emphasize that removing, keeping, and/or modifying a module influences the performance by means of mAP and fps. The dataset source also affects the outputs regarding mAP remarkably. For example, mAP decreases from 67.8% to 62.4% and from 54.4% to 52.3% for ReObj-MR and Re-Obj-GT, respectively, due to the removal of ICBs by IO-ReID-D from re-OBJ. On the other hand, adding full images compensates the drop in mAP and IO-ReID-DF becomes slightly better whilst exploiting BBXes instead of masks moreover contributes to mAP performance, where ReID accuracy is improved significantly without slowing down. *IO-ReID*consequently results in the fastest approach as it replaces ISB with YoloV3 without

encountering a considerable decrease in the mAP. As future work, we plan to further accelerate *IO-ReID* by knowledge sharing across instance segmentation and embedding generation modules.

Acknowledgment

Ertuğrul Bayraktar (EB) developed the idea, built the framework and conducted the experiments. EB also interpreted the results and wrote the paper.

References

[1] Suljagic H, Bayraktar E, Celebi N. Similarity based person re-identification for multi-object tracking using deep Siamese network. Neural Computing and Applications 2022; 34 (20): 18171-18182. https://doi.org/10.1007/s00521-022-07456-2

[2] Rubino C, Crocco M, Del Bue A. 3D Object Localisation from Multi-View Image Detections. IEEE Transactions On Pattern Analysis And Machine Intelligence 2017; 40 (6): 1281-1294. https://doi.org/10.1109/TPAMI.2017.2701373

[3] Nicholson L, Milford M, Sünderhauf N. QuadricSLAM: Dual Quadrics From Object Detections as Landmarks in Object-Oriented SLAM. IEEE Robotics And Automation Letters 2018; 4 (1): 1-8. **4**, 1-8 (2019). https://doi.org/10.1109/LRA.2018.2866205

[4] Zhu X, Zhu X, Li M, Moreiro P, Murino V et al. Intra-Camera Supervised Person Re-Identification. International Journal of Computer Vision 2019; 129 (5): 1580-1595. https://doi.org/10.1007/s11263-021-01440-4

[5] Bergmann P, Meinhardt T, Leal-Taixe L. Tracking without bells and whistles. In: Proceedings Of The IEEE International Conference On Computer Vision (ICCV); Seoul, Korea; 2019. pp. 941-951.

[6] Ye M, Shen J, Lin G, Xiang T, Shao L et al. Deep Learning for Person Re-identification: A Survey and Outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence 2021; 44 (6): 2872-93. https://doi.org/10.1109/TPAMI.2021.3054775

[7] Bansal V, James S, Del Bue A. re-OBJ: Jointly learning the foreground and background for object instance re-identification. In: Proceedings Of International Conference On Image Analysis And Processing (ICIAP); Trento, Italy; 2019. pp. 402-413.

[8] Bansal V, Foresti G, Martinel N. Where Did I See It? Object Instance Re-Identification with Attention. In: Proceedings Of The IEEE/CVF International Conference On Computer Vision; Virtual, Online; 2021. pp. 298-306.

[9] He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings Of The IEEE International Conference On Computer Vision; Venice, Italy; 2017. pp. 2961-2969.

[10] Gengec N, Eker O, Cevikalp H, Yazici A, Yavuz H. Visual object detection for autonomous transport vehicles in smart factories. Turkish Journal Of Electrical Engineering And Computer Sciences 2021; 29 (4): 2101-2115. http://doi.org/10.3906/elk-2008-62

[11] Özbek M, Syed M, Öksüz I. Subjective analysis of social distance monitoring using YOLOv3 architecture and crowd tracking system. Turkish Journal of Electrical Engineering and Computer Sciences 2021; 29 (2): 1157-1170. https://doi.org/10.3906/elk-2008-66

[12] Liang T, Chu X, Liu Y, Wang Y, Tang Z et al. Cbnet: A composite backbone network architecture for object detection. IEEE Transactions On Image Processing. 2022; 31: 6893 - 6906. https://doi.org/10.1109/TIP.2022.3216771

[13] Cheng G, Wang J, Li K, Xie X, Lang C et al. Anchor-free oriented proposal generator for object detection. IEEE Transactions On Geoscience And Remote Sensing. 2022; 60: 1-11. https://doi.org/10.1109/TGRS.2022.3183022

[14] Bian J, Mei X, Xue Y, Wu L, Ding Y. Efficient hierarchical temporal segmentation method for facial expression sequences. Turkish Journal Of Electrical Engineering And Computer Sciences 2019; 27 (3): 1680-1695. https://doi:10.3906/elk-1809-75

[15] Ayanzadeh A, Özuysal Ö, Okvur D, Önal S, Töreyin B et al. Improved cell segmentation using deep learning in label-free optical microscopy images. Turkish Journal Of Electrical Engineering And Computer Sciences 2021; 29 (8): 2855-2868. https://doi.org/10.3906/elk-2105-244

[16] Cheng B, Parkhi O, Kirillov A. Pointly-supervised instance segmentation. In: Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition; New Orleans, LA, USA; 2022. pp. 2617-2626.

[17] Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. In: Advances In Neural Information Processing Systems; Lake Tahoe, NV, USA; 2012. pp. 1097-1105.

[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations; San Diego, CA, USA; 2015. pp. 1-14.

[19] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S et al. Going deeper with convolutions. In: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (CVPR); Boston, MA, USA; 2015. pp. 1-9.

[20] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (CVPR); Las Vegas, NV, USA; 2016. pp. 770-778.

[21] Redmon J, Farhadi A. Yolov3: An incremental improvement. ArXiv TechReport ArXiv:1804.02767, 2018.

[22] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances In Neural Information Processing Systems; Montreal, Canada; 2015. pp. 91-99.

[23] Bolya D, Zhou C, Xiao F, Lee Y. YOLACT++: Better Real-time Instance Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022; 44 (2): 1108-1121. https://doi.org/10.1109/TPAMI.2020.3014297

[24] Gordo A, Almazán J, Revaud J, Larlus D. Deep image retrieval: Learning global representations for image search. In: Proceedings Of European Conference On Computer Vision (ECCV); Amsterdam, Netherlands; 2016. pp. 241-257.

[25] Radenović F, Tolias G, Chum O. Fine-tuning CNN image retrieval with no human annotation. IEEE Transactions On Pattern Analysis And Machine Intelligence. 2018; 41 (7): 1655-1668. https://doi.org/10.1109/TPAMI.2018.2846566

[26] Revaud J, Almazán J, Rezende R, Souza C. Learning with average precision: Training image retrieval with a listwise loss. In: Proceedings Of The IEEE International Conference On Computer Vision (ICCV); Seoul, Korea; 2019. pp. 5107-5116.

[27] Wei X, Luo J, Wu J, Zhou Z. Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Transactions On Image Processing 2017; 26 (6): 2868-2881. https://doi.org/10.1109/TIP.2017.2688133

[28] Jimenez A, Alvarez J, Nieto X. Class-weighted convolutional features for visual instance search. In: 28th British Machine Vision Conference (BMVC); London, UK; 2017. pp. 1-12.

[29] Salvador A, Nieto X, Marqués F, Satoh S. Faster r-cnn features for instance search. In: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition Workshops (CVPRW); Las Vegas, NV, USA; 2016. pp. 9-16.

[30] Teichmann M, Araujo A, Zhu M, Sim J. Detect-to-retrieve: Efficient regional aggregation for image search. In: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (CVPR); Long Beach, CA, USA; pp. 5109-5118.

[31] Le T, Nguyen K, Nguyen-Phan M, Ton T, Nguyen T et al. Others Instance re-identification flow for video object segmentation. In: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition Workshops (CVPRW); Honolulu, HI, USA; 2017. pp. 1-6.

[32] Tao R, Gavves E, Smeulders A. Siamese instance search for tracking. In: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (CVPR); Las Vegas, NV, USA; 2016. pp. 1420-1429.

[33] Li X, Qi Y, Wang Z, Chen K, Liu Z et al. Video object segmentation with re-identification.In: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (CVPR); Honolulu, HI, USA; 2017. pp. 1-6.

[34] Li X, Change-Loy C. Video object segmentation with joint re-identification and attention-aware mask propagation. In: Proceedings Of The European Conference On Computer Vision (ECCV); Munich, Germany; 2018. pp. 90-105.

[35] Farenzena M, Bazzani L, Perina A, Murino V, Cristani M. Person re-identification by symmetry-driven accumulation of local features. In: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (CVPR); San Francisco, CA, USA; 2010. pp. 2360-2367.

[36] Ahmed E, Jones M, Marks T. An improved deep learning architecture for person re-identification. In: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (CVPR); Boston, MA, USA; 2015. pp. 3908-3916.

[37] Paisitkriangkrai S, Shen C, Van Den Hengel A. Learning to rank in person re-identification with metric ensembles. In: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (CVPR); Boston, MA, USA; 2015. pp. 1846-1855.

[38] Bazzani L, Cristani M, Murino V. Symmetry-driven accumulation of local features for human characterization and re-identification. Computer Vision And Image Understanding 2013; 117 (2): 130-144. https://doi.org/10.1016/j.cviu.2012.10.008

[39] Bazzani L, Cristani M, Perina A, Murino V. Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recognition Letters 2012; 33 (7): 898-903. https://doi.org/10.1016/j.patrec.2011.11.016

[40] Kingma D, Ba J. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations; San Diego, CA, USA; 2015. pp. 1-15.

[41] Dai A, Chang A, Savva M, Halber M, Funkhouser T et al. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (CVPR); Honolulu, HI, USA; 2017. pp. 5828-5839.

[42] Viola R, Gautheron L, Habrard A, Sebban M. MetaAP: A meta-tree-based ranking algorithm optimizing the average precision from imbalanced data. Pattern Recognition Letters 2022; 161 (1): 161-167. https://doi.org/10.1016/j.patrec.2022.07.019

[43] Henderson P, Ferrari V. End-to-end training of object class detectors for mean average precision. In: Asian Conference On Computer Vision; Taipei, Taiwan; 2016. pp. 198-213.