Where Did I See It?
Object Instance Re-Identification with Attention

Vaibhav Bansal
vaibhav.bansal@uniud.it

Gian Luca Foresti
gianluca.foresti@uniud.it

Niki Martinel
niki.martinel@uniud.it

Machine Learning and Perception Lab, University of Udine, Italy

Abstract

Existing methods dealing with object instance re-identification (OIRe-ID) look for the best visual features match of a target object within a set of frames. Due to the nature of the problem, relying only on the visual appearance of object instances is likely to provide many false matches when there are multiple objects with similar appearance or multiple instances of same object class present in the scene. We focus on a rigid scene setup and to limit the negative effects of the aforementioned cases, we propose to exploit the background information. We believe that this would be particularly helpful in a rigid environment with a lot of reoccurring identical models of objects since it would provide rich context information. We introduce an attention-based mechanism to the existing Mask R-CNN architecture such that we learn to encode the important and distinct information in the background jointly with the foreground features relevant to rigid real-world scenarios. To evaluate the proposed approach, we run compelling experiments on the ScanNet dataset. Results demonstrate that we outperform significantly compared to different baselines and SOTA methods.

1. Introduction

In the field of computer vision, multiple object matching and association are two of the classical problems that find applications in many distributed systems tackling video surveillance, semantic scene understanding, and Simultaneous Localization And Mapping (SLAM) tasks among others. Such tasks are particularly challenging in indoor environments where the scenes are generally cluttered with many objects that makes it difficult to correctly identify and track a specific object instance among a set of almost-identical ones (see Figure 1 for a few samples). The problem is further compounded by the common wide baselines shared among different/temporally disjointed views. Relying on the visual appearance to re-identify object instances

Figure 1. Few samples from the ScanNet dataset [6]. Different similar-looking object instances are difficult to differentiate with each other. However, the context information provided by the background can play a relevant role to ease the re-identification of a particular instance in multiple views.

in different images is thus a very complex task facing a variety of challenges for the association problem, i.e.: occlusions, motion blur, misdetections, etc. To address such challenges, existing re-identification methods can be categorized into two major groups of approaches: appearance-based and motion-based. Appearance-based approaches are the most widely investigated ones because motion-based systems suffer from spatio-temporal constraints. Such methods try to localize each object instance based on a motion model, that, due to the possibility of severe and uncontrollable trajectories through the frames, tends to fail when the same object instance is perceived after a long time.

A closely related task to the object instance re-identification is the *person* re-identification problem where the goal is to re-associate the image of a query person among a set of (gallery) instances already acquired by multiple disjoint cameras at different time instants. Existing person re-identification methods try to learn discriminative features based on person's appearance (e.g., [16, 11]). The problem of associating a unique ID to instances of objects is however different since the aim is to associate multiple unknown objects between different (overlapping) views [18]. A closely related previous work, re-OBJ [1] recognized the

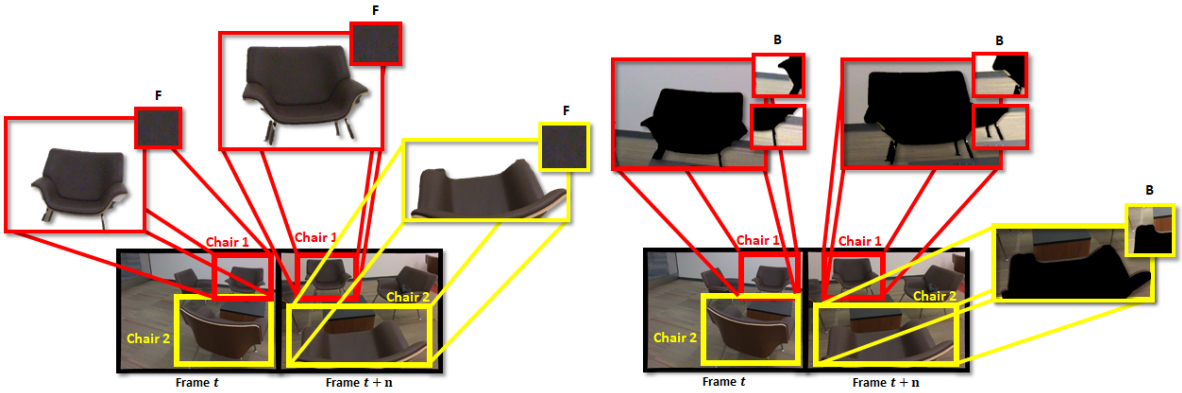

Figure 2. An example scene with a collection of chairs. (Left) If only foreground appearance is considered, they are identical chairs across multiple views, marked as *F* in insets. Any affine-invariant descriptor would easily match the incorrect chairs. (Right) However, if we also consider the background there is useful discriminative information such as other objects, the wall or the skirting board, marked as *B* in insets.

challenge of re-identification of multiple object instances in multiple views in an indoor scenario for the first time. re-OBJ [1] proposed to jointly encode the foreground appearance of different object instances along with partial observations of the background for an object instance re-identification framework suited for a rigid environment.

Most of the previous attempts for object re-identification utilize CNN-based neural networks. CNNs rely on a couple of implicit inductive biases that help them to generalize well on lesser data [10]: 1) Locality: the neighbouring pixels in an image are related. CNNs use sliding filters over a small patch of an image to exploit local dependencies. 2) Weight Sharing: different parts of the image are processed in an identical fashion regardless of their absolute location. Moreover, the downsampling operations such as pooling and strided convolutions further compress the information loosing a lot of relevant features in the process. Although, CNNs have proved to be quite successful in general classification tasks, they are unable to make relationship among distant-pixels and such lack of crucial spatial information make them not so suitable for tasks like object instance re-identification. Furthermore, most of the existing algorithms in the literature rely on building a model of the target objects by learning the appearance of only the target objects in the foreground and by applying the learned appearance model to match the target objects across multiple views. But, re-OBJ [1] shows that if only the objects in the foreground are considered, then any affine-invariant descriptor would learn to find similarities between two incorrect instances because they look very similar as only learnt from their foreground appearances (see Figure 2). However, if we consider a static indoor scenario where large displacement in the camera motion is unlikely and so the background of an instance cannot undergo a sudden drastic change, then the background can provide highly discriminative information. Consider, two different views of the same static scene where the objects are stationary and only the camera is moving, the background might contain extremely important cues such as border between the wall and the floor, other objects in vicinity and other useful discriminative features present in the scene (see Figure 3).

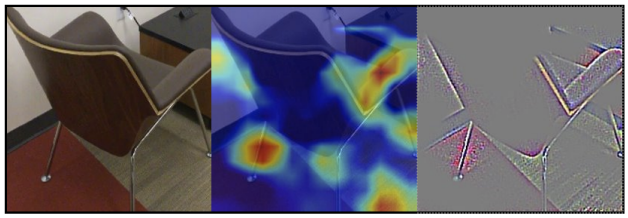

Figure 3. Left: The image of an object instance in a static scene. Centre: The image of the instance viewed through Grad-CAM [22] highlighting gradient weighted high importance regions. Right: Guided Grad-CAM image to show the pixel-space gradient visualizations to better understand the discriminative features present in the background.

Taking inspiration from [1], we propose to use transformer-based models such as ViT [9]) to harness the long-range dependencies from the background information in addition to the foreground appearance in order to discriminate among multiple instances of the same semantic class and also among the objects that have a similar appearance as shown intuitively in Figures 1 and 2. To include the background information, we use an off-the-shelf object detector, i.e. Mask-RCNN (sec. 3.1), and obtain foreground & background masks of the objects with the bounding boxes that are expanded (see sec. 4) to include a substantial background around the object within the bounding boxes. Then, to capture and encode the spatial and context information around each selected region, we leverage the recent concepts behind the transformer-based models

(e.g., ViT [9]) instead of CNN-based models. The architecture of self-attention-based models allows them to have minimal inductive biases. ViT interprets an image as a sequence of patches and, to process such a sequence, it uses a self-attention mechanism that models the relationships between its elements. The multi-headed self-attention allows ViT to attend to all the patches of the sequence in parallel and harness long-range dependencies between several patches across different frames. Precisely, foregrounds and the masked backgrounds are fed to a pre-trained ViT model [9] (sec. 3.2) to extract the region encodings, which are then further processed by a self-attention mechanism that will separately embed foreground and background information into query, key and value to better exploit the interaction between the foreground and background context. The resulting embedding is then considered into a triplet-based network architecture (see Figure 6) with the pairwise ranking model to learn similarity at the instance level for a triple-based ranking loss function.

Thorough experiments conducted on the ScanNet dataset [6] demonstrate that our method significantly outperforms existing approaches and yields to substantial improvements over different baselines.

2. Related Work

The object re-identification task is well-studied in the literature. However, it has mostly been presented as a person re-identification problem. The object re-identification task aims to re-identify objects in the images by using visual search to retrieve a similar set of images for a given query image of the target object. Earlier models [4, 5, 23] in the literature simply extract features like Gabor filters, SIFT [17], HOG [7] features to learn image similarity. However, the representations using the hand-crafted features were limiting the performance of such methods. Some deep learning-based models successful in image classification [15] have been used to learn features from the images but these models could not retrieve the fine-grained distinction between similar images. [24] proposed a pairwise ranking model in order to learn fine-grained image similarity. Pairwise ranking model was successfully used to learn image ranking models in [5, 12, 19]. FaceNet [20] used a similar ranking model but replaced verification loss [21] with triplet loss for the verification, recognition and clustering.

Most of the previous studies like [24, 20, 21, 3, 11] rely on learning an appearance-based transfer function for a robust re-identification system. Moreover, [11] extracted features from three different modalities such as the chromatic content, spatial arrangement of colors and local motifs derived from different parts of the human body to accumulate local features. In [19], the authors trained a model to rank images based on the relative attributes among the images with the similar attributes. OASIS [5] computes local distance [12] to learn an image similarity ranking model in addition to the hand-crafted features. However, these appearance-based methods are good at identifying only the intra-class variation, they usually fail to perform well in identifying recurring multiple instances of the same object class in different views. The applications that involve computation of image similarity like re-identification, image retrieval, search-by-example require learning a fine-grained image similarity that can also distinguish the differences between different images of the same category. Thus, we focus on the objects' relationship to the background in order to learn a unique discriminative feature for a particular object instance. We take inspiration from a closely related work [1] that encodes the foreground appearance and partial observations of the background using a pre-trained ResNet [14] for identifying multiple instances of objects with same semantic class in different views. Instead of a CNN-based architecture like ResNet, we propose to use ViT [9] to generate the joint features which are then further processed by an additional self-attention block which finds long-range dependencies between different image patches and maintains the vital spatial resolution at the output to learn a unique discriminative representation for each object instance.

3. OIRe-ID Framework

3.1. Object Detection

Our method builds upon any off-the-shelf object detection algorithm such as Mask-RCNN [13] which uses region-based object detector like Faster R-CNN to detect objects. Mask-RCNN not only provides a bounding box around an object but also performs image segmentation and provides a mask representing a set of pixels belonging to the same object. We use Mask-RCNN to extract bounding boxes including masks as separate images and resize them into images of a fixed size in order to train our network to learn a visual encoding of the objects' mask and the background surrounding them within the bounding boxes (see Figure 4).

3.2. Foreground and Background Joint Feature

For each object of the input images, we create two sets of images $F = \{I_f, I_b\}$. Using the detections obtained from Mask-RCNN, one set is created by extracting masks representing objects in the foreground (I_f). The other set only contains the background with the subtracted foreground (I_b). As shown in Figure 4, a pair of images is taken from each set to pass through two identical streams to learn an encoding for the masked foreground and the background. Each of the images, the masked background and the masked foreground is input to a ViT [9] model pre-trained on ImageNet [8] dataset.

An input image I to ViT is divided into N fixed-size

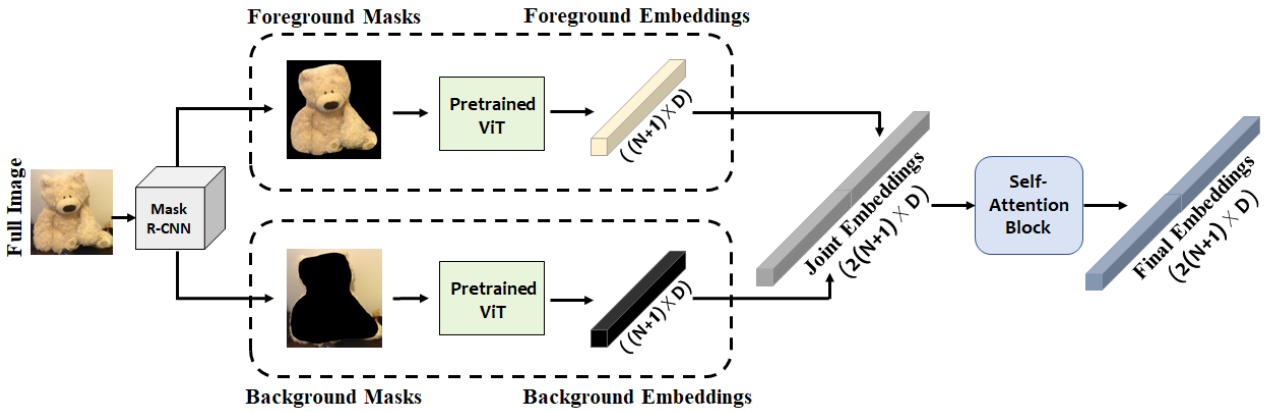

Figure 4. Our network takes expanded bounding boxes (see sec. 3.2)containing masked foreground and masked background of different object instances. Each masked pair is fed to a ViT [9] network to generate a embeddings of dimensions $(N+1) \times D$, which are then concatenated to provide a joint representation of $2(N+1) \times D$ before passed down to a self-attention block.

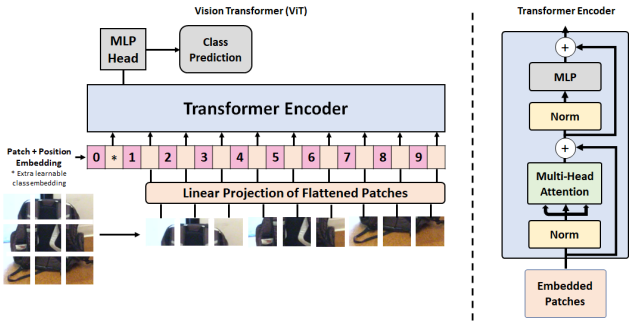

Figure 5. Architecture of ViT [9]. An input image is first divided into equal size patches to form a sequence. Then, a learnable position embedding is added for spatial information and an extra learnable *cls* token is added before the input sequence to the encoder.

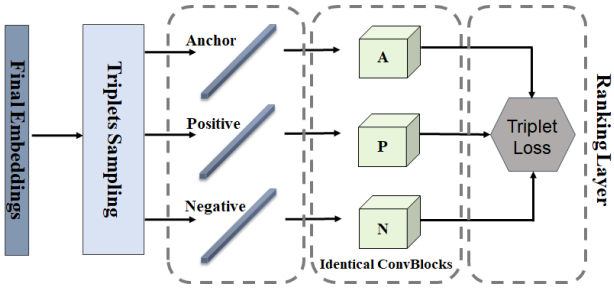

Figure 6. Triplet sampling of the joint embeddings. The joint embeddings are segregated into embeddings of the anchor image A, positive image P and a negative Image N which are fed into three identical networks for optimizing the triplet loss.

patches $(I^t{}_p | t = 1, 2, ..., N)$. An extra learnable *cls* embedding token denoted as I_{cls} is pre-pended to the input sequences (See Figure 5). The embeddings of dimension $\mathbb{R}^{(N+1)\times D}$ obtained as an ouput from ViT are representation of the two images retaining spatial and semantic context. These embeddings are then concatenated to provide a joint embedding and further processed by a self-attention block giving a final representation $\mathbb{R}^{2(N+1)\times D}$.

3.3. Triplet Loss for Instance Re-identification

To effectively handle object instance re-identification, an optimal system should be able to distinguish not only among the images of different objects but also among different instances of objects of the same semantic class. Especially considering the rigid and static indoor scenario where multiple instances of the same object category are present, it is highly challenging to re-identify a particular object instance amongst others. As shown in Figure 6, the final embeddings are arranged into a triplet of images for passing down to a triplet-based training architecture.

A triplet constitutes three kinds of images: an *anchor* which is the query template, a *positive* and a *negative* image. A *positive* image is simply a transformed *anchor* image. A *negative* image has to be carefully selected for an effective re-identification at the instance level.

For instance, an image pair of two different classes of objects (say, a cup and a table) is definitely an example of *anchor-negative* pair but two different instances of the same object (say, two different cups on the table) is also considered an *anchor-negative* pair. We use a triplet-based network architecture with the pairwise ranking model to learn image similarity for the triple-based ranking loss function, inspired from [24]. Consider, a set of $F = f_1, ...f_F$ images and $s_{i,j} = s(f_i, f_j)$ representing the pairwise similarity score between the images f_i and f_j. The score s is higher for more similar images and is lower for more dissimilar images. If we have a triplet $t_i = (f_{iA}, f_{iP}, f_{iN})$ where f_{iA}, f_{iP} and f_{iN} are the anchor, positive and negative images, respectively, then the training goal is to learn

an embedding function such that:

$$D(f_{iA}, f_{iP}) < D(f_{iA}, f_{iN}), s(f_{iA}, f_{iP}) > s(f_{iA}, f_{iN}) \quad (1)$$

where $D(.)$ is the squared Euclidean distance in the embeddings space. A triplet incorporates a relative ranking based on the similarity between the anchor, positive and the negative images.

The triplet ranking loss function is given as:

$$l(f_{iA}, f_{iP}, f_{iN}) = max\{0, M + D(f_{iA}, f_{iP}) - D(f_{iA}, f_{iN})\} \quad (2)$$

where M is a parameter called *margin* that regulates the gap between the pairwise distance: (f_{iA}, f_{iP}) and (f_{iA}, f_{iN}). The model learns to minimize the distance between more similar images and maximize the distance between the dissimilar ones.

4. Experiments

Training data. To evaluate the performance of our proposed method, we use a video dataset, ScanNet [6] for our experiments which consists of 1500 indoor RGBD scans annotated with 3D camera poses, surface reconstructions, and mesh segmentation related to several object categories. To generate our training data, we employ Mask-RCNN over a subset of 863 scenes randomly selected from the whole ScanNet dataset. Overall, Mask-RCNN generated $646,156$ object detections with masks belonging to 29 object classes (see Table 1). We use the ground-truth available with the dataset to evaluate the accuracy of Mask-RCNN on the ScanNet images. We discarded the detections for which no ground-truth annotations were available. We estimated bounding box overlap ratio between the ground truth (GT) bounding boxes and the Mask-RCNN detections to collect *valid* detections. The overlap ratio was chosen to be higher than 60% and the label of the detected object should match with the GT label for any detection to be considered a *valid* detection.

Finally, we found around 9.11% of the total detections (around 58876) to be considered *valid* for the experiments. The regions contained by the bounding boxes were stretched by an additional 10 pixels in all directions to obtain loosely-fitted bounding boxes around the objects to allow larger background information around each object's foreground mask within the bounding boxes. These regions are resized to 224×224 and stored according to the object's class and it's observed instances in a particular scene. At the end, the foreground masks and the background masks were extracted from these regions and stored as separate images for each object category. The data is split into a 3-fold cross-validation manner with 39250 images for training and 19626 images for test over 1701 instances of objects.

Table 1. Category-wise number of views and unique instances obtained from ScanNet after filtering out only the *valid* detections selected based on object's label and the bounding box overlap ratio with the ground-truth [1]

No. of Views and Unique Instances Per Object Class					
Class	No. of Views	No. of Instances	Class	No. of Views	No. of Instances
bicycle	110	6	toilet	1755	103
bench	27	4	tv	562	46
backpack	1563	117	laptop	600	41
handbag	486	32	mouse	59	6
suitcase	377	30	keyboard	1879	67
sports ball	379	21	microwave	667	61
bottle	903	27	oven	72	6
cup	278	25	toaster	11	4
chair	38203	508	sink	2694	157
couch	1371	75	refrigerator	60	11
potted plant	1294	55	book	3124	65
bed	83	17	clock	25	6
bowl	121	8	person	260	8
dining table	1853	185	teddy bear	47	8
vase	13	2	-	-	-

We performed experiments in three different setups as in [1]. In all the experimental setups, we used pre-trained ViT [9] on the ImageNet [8] dataset as the feature extractor backbone. **no-train:** In this configuration, the full image regions were matched directly against each other by using an $L2$ distance-based metric, without any training. **full:** In this configuration, our proposed model is trained using the full images without extracting separate foreground and background masks. **concat:** In this experimental setup, the approach proposed in this paper is used by training the model with the concatenated embeddings obtained from masked foregrounds and the backgrounds. In *concat* setup, the model undergoes a triplet-based training which learns to minimize the difference between the anchor f_{iA} and the positive f_{iP} images while maximizing the difference between the anchor f_{iA} and the negative f_{iN} images at the same time.

Evaluation Metrics. We use the standard Cumulative Matching Characteristic (CMC) and Rank-1 accuracy as the metric to evaluate the performance of our method against other existing methods. The performance of the framework is judged by the number of good matches of the probe image with a gallery of images in rank-1.

Table 2. Ablation study with different experimental setups to evaluate our method on the ScanNet with Rank-1,-5,-20 and -50 accuracy values. *concat* depicts the proposed approach.

type	**Rank-1**(%)	**Rank-5**(%)	**Rank-20**(%)	**Rank-50**(%)
no-train	68.7	77.06	81.78	92.71
full	75.79	88	91.22	96.25
concat	**83.89**	**94.61**	**99.42**	**100**

Analysis Table 2 shows the performance based on the three different experimental setups on the ScanNet dataset in regards to rank-1 accuracy. The results show that the proposed method represented as *concat* in the table was able to improve the rank-1 accuracy by 15.19% and 8.1% against *no-train* and *full*, respectively.

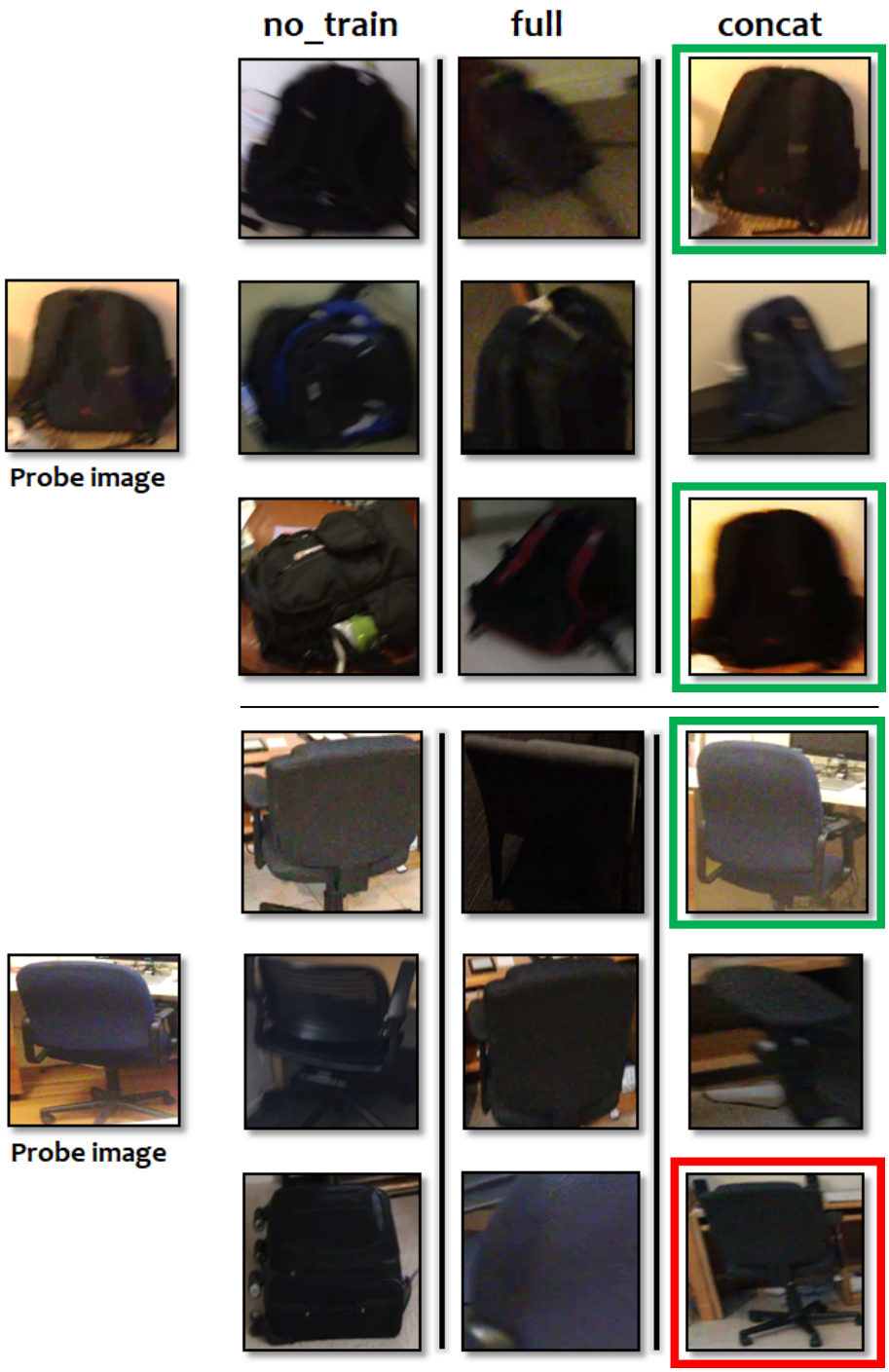

Figure 7. Visualizing the matches found in *no-train*, *full* and *concat* experimental setups. The right matches with the probe image are highlighted in green color. The red bounding box highlight a match that contains a different instance of the same object class. We can observe that in such indoor scenarios, the background information can be highly useful in order to discriminate among different instances of the same object.

Figure 7 shows that the proposed method, *concat* was able to find the best match with the probe image. Other setups, *no-train* and *full* tried to match with an image where the object has either same color or the shape. However, the proposed method, *concat* could not only handle the intra-class variations but could also distinguish among different

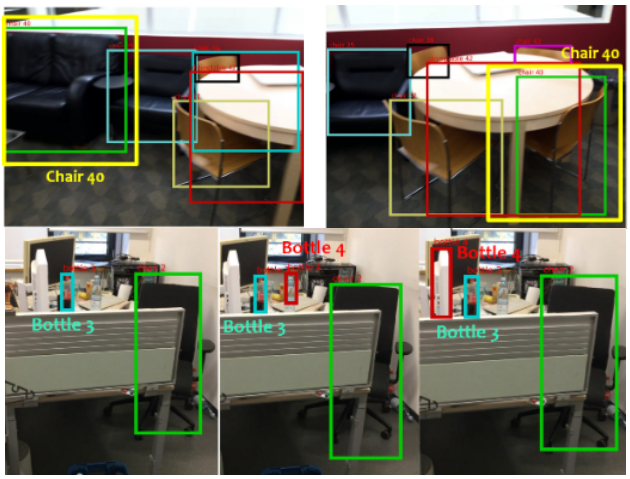

Figure 8. Examples of challenges faced by standard object tracking and association algorithms like deepSort in a cluttered environment. (Top) ID switching for the same object instance (Chair 40 highlighted in yellow box) and, (Bottom) When camera revisits the same region later, same object instance is assigned a different ID (Bottle 4 highlighted in red box)

instances of the objects with same semantic class. However, occasionally our method could not identify the exact instance but found a match with another instance of the same object category as highlighted in red bounding box.

Comparison with state-of-the-art To evaluate our approach with the state-of-the-art methods, we compare our method with the previous works, re-OBJ [1] and deepSort [25] which is a multi-object tracking algorithm repurposed here as a rank-1 re-ID method. deepSort [25] is based on SORT [2] algorithm that utilizes deep appearance descriptors for better accuracy in multiple object tracking. deepSort employs a deep association metric by learning discriminative feature embeddings offline on a pedestrian dataset. For our evaluation, we fed two random pairs of images obtained from the ScanNet scenes to deepSort to associate multiple objects. We evaluated the performance by measuring the percentage of matched object instances across all the image pairs. Figure 8 shows the kind of challenges that standard object matching or tracking algorithms face in re-identifying objects in cluttered indoor scenes.

While the deepSort was able to match a few objects (*bottle* in blue bounding box) in multiple frames but lost many objects such as chair (in yellow bounding box) and another bottle (in red bounding box), especially, when the camera revisits the same region of the scene at a different point of time. Since such environments are cluttered with several objects and there are multiple instances of similar looking objects are present, state-of-the-art object matching and association algorithms fail to perform well.

Table 3 shows the comparison in performance of our method to multiple pedestrian tracking algorithm, deepSort and the state-of-the-art CNN-based object instance re-identification method, re-OBJ. deepSort and re-OBJ achieved a rank-1 accuracy of 49.60% and 77.85%, respectively, against the rank-1 accuracy of 83.89% obtained with our method.

Table 3. Rank-1 accuracy of our method in comparison to state-of-the-art multiple object tracking algorithm, deepSort [25] and CNN-based re-OBJ [1]

method	**Rank-1**(%)
deepSort	49.60
re-OBJ	77.85
ours	**83.89**

4.1. Experiments with person re-ID Baseline

To compare the performance of our method with other existing state-of-the-art methods in object re-ID, we performed some experiments with the methods for person re-identification methods.

OSNet We used one such baseline re-ID method like Omni-Scale Feature Learning for Person Re-Identification (OSNet) [27] on our data. OSNet tries to address person re-identification (ReID) as an instance-level recognition problem by not only capturing different spatial scales but also encapsulating a combination of multiple scales. The authors define a combination of features of both homogeneous and heterogeneous scales as omni-scale features. For example, features of variable homogeneous scale while identifying a person in multiple cameras would be a combination of global-scale features like the whole body region including the clothes and the corresponding local-scale features like the shoes, glasses etc. But, these features could be shared among different individuals like two different people wearing the similar clothes and shoes. Thus, apart from these features of variable homogeneous scales, more complicated and richer features would be required. For example, when two people are wearing the clothes of same color (say, a white shirt), a specific logo in the front could be the distinguishing factor between the two. Although, the logo alone won't be distinctive enough on its own without the consideration of the clothes as a context which might otherwise be confused with several other patterns in the scene. Thus, the unique combination of small-scale features like the logo size on the shirt and the medium scale features like the upper-body size of the person could constitute the features of variable heterogeneous scales. In this work, a deep CNN-based ReID is proposed for learning these so called omni-scale features. This is achieved by designing a residual block composed of multiple convolutional streams, each detecting features at a certain scale. Importantly, a novel unified aggregation gate is introduced to dynamically fuse multi-scale features with input-dependent channel-wise weights.

Table 4. Performance of our method in comparison to person re-ID methods. The best performing type of setups is highlighted in bold.

method	**Rank-1**(%)	**Rank-5**(%)	**Rank-10**(%)	**Rank-20**(%)
OSNet	69	85.7	89	91.3
DGNet	58.3	76	83.7	92.4
re-OBJ	77.85	91.55	-	98.36
ours	**83.89**	**94.61**	**-**	**99.42**

DGNet Another baseline method we used for our experiments is Joint Discriminative and Generative Learning for Person Re-identification (DGNet) [26]. DGNet proposes a joint framework to couple the discriminative and generative learning to improve the learned re-id embeddings by leveraging the generated data. The generative module separately encodes each target (person) into an appearance code and a structure code, and a discriminative module shares the appearance encoder with the generative module. The appearance space encodes appearance of the person and other related semantic features including color of clothing, shoes, texture and style etc. while the structure space encode geometrical and spatial information including body size, volume of hair, pose, background and viewpoint, etc. By switching the appearance or structure codes, the generative module is able to generate high-quality cross-id composed images, which are then fed back online to the appearance encoder which is further used to improve the discriminative module. The results in comparison to the proposed approach are given in Table 4.

5. Conclusion

The contribution of this paper was to explore the intuition that the information obtained from the background surrounding the detected target objects in a rigid scene could be highly useful in discriminating two near-identical objects or two instances of the same object class. The discriminative features learned from the explicit concatenated foreground and background can be utilized to re-identify objects at the instance-level throughout the dataset. Our experiments have shown that the proposed method based on self-attention-based transformer model performs well even in the case of highly cluttered rigid environments like the indoor scenes obtained from ScanNet dataset. In future, we plan to explore if the temporal information obtained from multiple views in a video dataset can be integrated with our object instance re-identification system for a robust multiple object tracking algorithm in case of rigid and static scenes.

6. Acknowledgment

This work was partially supported by ONR grant N62909-20-1-2075.

References

[1] Vaibhav Bansal, Stuart James, and Alessio Del Bue. re-obj: Jointly learning the foreground and background for object instance re-identification. In *International Conference on Image Analysis and Processing*, pages 402–413. Springer, 2019.

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

[3] A. Bhuiyan, A. Perina, and V. Murino. Exploiting multiple detections to learn robust brightness transfer functions in re-identification systems. In *Proc. IEEE International Conference on Image Processing*, pages 2329–2333, Sep. 2015.

[4] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2559–2566, June 2010.

[5] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(Mar):1109–1135, 2010.

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE, 2005.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] FacebookAI. Better computer vision models by combining transformers and convolutional neural networks, 2020.

[11] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2360–2367. IEEE, 2010.

[12] Andrea Frome, Yoram Singer, and Jitendra Malik. Image retrieval and classification using local distance functions. In *Advances in Neural Information Processing Systems*, pages 417–424, 2007.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE*

conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[16] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan. Clothing attributes assisted person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):869–878, May 2015.

[17] David G Lowe. Object recognition from local scale-invariant features. In *Proc. IEEE International Conference on Computer vision*, volume 2, pages 1150–1157. IEEE, 1999.

[18] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[19] Devi Parikh and Kristen Grauman. Relative attributes. In *Proc. International Conference on Computer Vision*, pages 503–510. IEEE, 2011.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[21] Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. In *Advances in Neural information processing systems*, pages 41–48, 2004.

[22] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[23] Graham W Taylor, Ian Spiro, Christoph Bregler, and Rob Fergus. Learning invariance through imitation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2729–2736. IEEE, 2011.

[24] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.

[25] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *Proc. IEEE International Conference on Image Processing*, pages 3645–3649. IEEE, 2017.

[26] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2138–2147, 2019.

[27] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. *CoRR*, abs/1905.00953, 2019.