Object ReID in an office environment: An empirical study

Dmytro Klepachevskyi
University of Waterloo
Vision and Image Processing Lab, Critical ML Lab
dklepach@uwaterloo.ca

Sirisha Rambhatla
University of Waterloo
Critical ML Lab
sirisha.rambhatla@uwaterloo.ca

Yuhao Chen
University of Waterloo
Vision and Image Processing Lab
yuhao.chen1@uwaterloo.ca

Abstract

Object Re-identification (ReID) is a fundamental task in computer vision, enabling systems to recognize and track the same object across different frames and viewpoints, lighting conditions, and environmental contexts. In robotic applications, reliable object ReID is essential for enabling robots to maintain persistent identity of objects over time. While person and vehicle ReID have been extensively studied, object-level ReID remains unexplored. In this work, we present an empirical comparative study of state-of-the-art representation learning algorithms - DINO, DINOv2, Triplet, I-JEPA, and CLIP, which are applied to object ReID in an office environment. We construct a custom office dataset, capturing diverse office objects. Each image is cropped using Grounding DINO for object detection. We extract embeddings for each object instance and perform ReID by computing cosine similarity. Performance is assessed by measuring whether the top-matching image corresponds to the correct object, using Mean Average Precision, Top-1 and Top-5 metrics.

1. Introduction

Object Re-Identification (ReID) aims to detect and track specific objects across various camera viewpoints and environments. It plays a crucial role in surveillance systems and autonomous robotic applications, where the goal is to have a continuous mapping between detections and unique objects known as an open-world re-identification [11]. This leads a foundation in robotic systems, where a robot is able to re-identify the object and track it, or bring it to a person. Current solutions often focus on person ReID [12, 15, 16, 18], vehicles ReID [1, 10], or pet reID [14], therefore, there is a lack of studies on generalization to various objects. These methods remain dependent to a specific category, and they struggle to generalize on other unseen objects. Object representation learning methods are powerful for ReID tasks and usually used for person, vehicle, and pet ReID [17].

Current object representation methods are dominated by self-supervised learning (SSL) techniques, which aim to learn discriminative embeddings without requiring manual labels. These approaches typically maximize the similarity between different augmented views of the same image while minimizing similarity between views of different images, using a contrastive learning objective. Methods using this paradigm of learning include DINO [3], DINOv2 [7], CLIP [8], Triplet loss learning [5], and I-JEPA [2]. Although all of them produce high-quality embedding spaces, they differ in training strategy and the type of supervision which is used. However, these methods were extensively tested in person, vehicle, and animal ReID tasks, and still lack evaluations with various objects.

To enable generalization to open-set vocabularies of arbitrary objects, it is essential to evaluate existing ReID approaches beyond traditional domains such as person, vehicle, and animal re-identification. Indoor environments like offices which contain various objects (e.g., monitors, keyboards, books, personal items) whose identification may need to be consistently recognized across viewpoints. Studying ReID performance in such settings is therefore crucial for understanding the limits of current methods and their ability to generalize to broader object categories.

This paper presents a comparative analysis of the SSL methods used in an office environment for various object ReID, with the goal to identify potential limitations and future improvements. Our primary contributions include:

1. A pipeline for extracting cropped object images and constructing a database with assigned IDs, using Grounding DINO [6] for detection and Segment Anything Model 2

(SAM 2) [9] for precise segmentation across frames.

2. Extraction of feature embeddings using five models, followed by a ReID evaluation using an input query image and three metrics: Mean Average Precision (mAP), Top-1 and Top-5 accuracies.

2. Related Work

2.1. Models overview

Vision foundation models such as DINO and DINOv2 utilize large-scale self-supervised learning to obtain general-purpose visual representations that transfer effectively across domains. In this study, we evaluate five representative embedding models - DINO, DINOv2, CLIP,

Triplet learning, and I-JEPA, whereas each model learns similarity-preserving features.

DINO employs a student–teacher self-distillation framework in which the student network is trained to match the teacher's predictions across augmented image views.

DINOv2, an improved successor of DINO, combines curated large-scale training data with architectural refinements and more stable optimization strategies.

CLIP adopts a multimodal contrastive objective that aligns images with their textual descriptions, allowing it to learn rich semantic embeddings but with behavior that differs from instance-level discrimination.

DINO, DINOv2 and CLIP rely on a contrastive learning objective that encourages embeddings of different augmented views of the same image (positives) to be similar, while pushing apart embeddings of other images (negatives). A standard InfoNCE contrastive loss (Eq. 1) is used.

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp\left(\text{sim}(z_i, z_i^+)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(z_i, z_j)/\tau\right)}, \quad (1)$$

where z_i and z_i^+ denote embeddings of two augmented views of the same image, $\text{sim}(\cdot)$ is cosine similarity, and τ is a temperature parameter.

Triplet learning is based on supervised metric learning, optimizing embeddings such that an anchor is closer to a positive sample (same identity) and farther from a negative sample.

Triplet-based model optimizes a supervised metric learning objective (Eq. 2) that enforces an anchor a to be closer to a positive sample p (same identity) than to a negative sample n (different identity).

$$\mathcal{L}_{\text{triplet}} = \max\left(0,\ d(f(a), f(p)) - d(f(a), f(n)) + \alpha\right), \quad (2)$$

where $d(\cdot)$ is a distance function (e.g., Euclidean) and α is the margin.

Finally, I-JEPA introduces a non-contrastive predictive objective, where the model predicts high-level representations of masked spatial contexts.

I-JEPA uses a non-contrastive predictive objective (Eq. 3). Instead of comparing augmented views, the model predicts the latent representation of masked target regions using the context region.

$$\mathcal{L}_{\text{I-JEPA}} = \|g_\theta(x_{\text{context}}) - f_\phi(x_{\text{target}})\|_2^2, \quad (3)$$

where g_θ predicts high-level representations of masked spatial regions, and f_ϕ encodes the ground-truth target regions.

2.2. Similarity measurement

Cosine similarity (Eq. 4) is used to measure the closeness between feature embeddings extracted from the query image and those stored in the cropped object database.

$$\cos(x, y) = \frac{x \cdot y}{\|x\|\,\|y\|}. \quad (4)$$

Given two embedding vectors, cosine similarity measures how aligned they are in the high-dimensional feature space, independent of their magnitude. This makes it particularly suitable for representation learning methods, where the direction of the embedding vector captures semantic information.

2.3. Evaluation Metrics

We evaluate the re-identification performance using three standard metrics: mean Average Precision (mAP), Top-1 accuracy, and Top-5 accuracy.

Top-k accuracy measures whether the correct identity appears within the top k highest-ranked database images based on cosine similarity. Top-1 accuracy indicates whether the top retrieved match is correct, while Top-5 accuracy verifies whether the correct identity is found among the five most similar results.

mAP provides a more comprehensive retrieval evaluation. For each query, the Average Precision (AP) measures how consistently relevant images are ranked above irrelevant ones. mAP (Eq. 5) is then computed as the mean AP across all queries:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{i=N} \text{AP}_i, \quad (5)$$

where N is the number of query images.

Together, these metrics provide insights into the retrieval behavior of the evaluated models.

3. Methodology

As shown in Figure 1, the pipeline workflow consists of three main stages: dataset preprocessing, object re-identification using pre-trained models to extract feature embeddings, and similarity computation between a new input image and the feature vectors stored in the cropped object database.

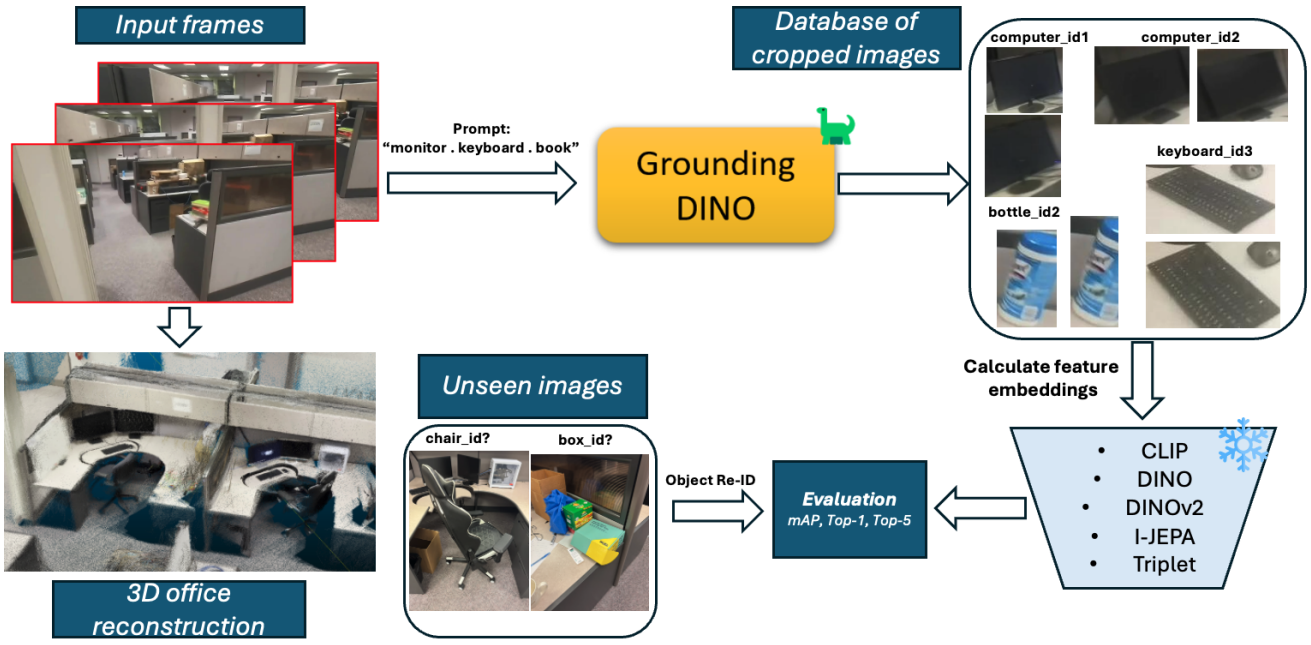

Figure 1. Overview of the proposed object ReID pipeline. Input frames are processed with GroundingDINO via prompt with objects. Each distinct object is assigned with a unique ID. For each image, the feature representation embeddings are calculated with 5 methods and evaluated with mAP, Top-1, Top-5 metrics.

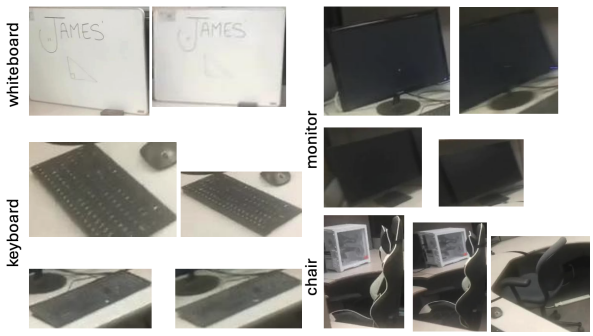

Figure 2. Overview of the cropped dataset. Each input image is processed with Grounding DINO to retrieve an object and crop it. Each object can have multiple cropped images captured from different viewpoints.

3.1. Dataset Acquisition and Pre-processing

The dataset (Figure 2) was collected in the office environment with total of 521 frames, which were used to get cropped images for object ReID. To adapt the dataset for the ReID task, we utilized GroundingDINO for object detection via textual prompts, and SAM 2 for precise segmentation over frames. The correspondings IDs of each distinct object are preserved for each cropped image.

3.2. Feature embeddings calculations

Once we have cropped images of different classes and preserved IDs, the embeddings are calculated using 5 previously described models: DINO, DINOv2, CLIP, Triplet, I-JEPA. Finally, for each cropped image's embedding we find the most closest ones in the database, and measure the accuracy using mAP, Top-1 and Top-5 accuracies.

An important note is that a derived class from GroundingDINO prompt is compared only with the same class, which might lead to better accuracy.

3.3. Evaluation on Previously Unseen Data

In addition to evaluating performance on the full dataset, we conduct an additional experiment using previously unseen images. For this experiment, a new set of images was manually captured using a phone camera. Grounding DINO was again applied to these images using textual prompts corresponding to the target object categories to obtain cropped query objects. Once a query crop was extracted, its feature embedding was computed using each of the five models.

The embedding of the query was then compared with all stored embeddings in the cropped-object database, and the system retrieved the top three most similar objects based on cosine similarity. This setup simulates a realistic use-case in which a user presents a new photo and the system attempts to identify the same object instance from the database.

Table 1. Comparison of re-identification performance of various models on the office dataset.

	mAP	Top-1	Top-5
CLIP [8]	60.8	62.5	79.3
DINO [3]	76.4	69.3	89.1
DINOv2 [7]	80.4	76.3	92.3
I-JEPA [2]	66.7	65.9	88.2
Triplet [5]	**85.1**	**79.1**	**94.4**

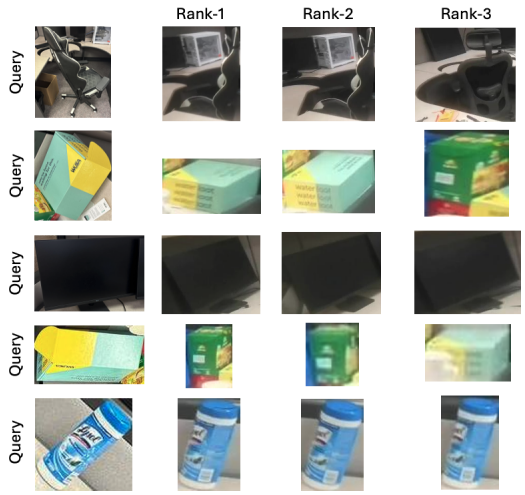

Figure 3. Ranking results of the best-performing model (Triplet) on previously unseen query images. The queries include a chair, a box, a monitor, a flipped box, and a rotated bottle.

4. Experiments and Results

In the experimentation setup, we retrieve cropped images of 11 next popular objects in the office environment: monitor, keyboard, mouse, box, bottle, cabel, whiteboard, chair, cup, marker, book.

The quantitative results are described in the Table 1, and highlight, that the Triplet algorithm obtained the best results, achieving 85.1% of mAP, 79.1% of Top-1 accuracy, and 94.4% of Top-5 accuracy. All the algorithms achieved high Top-5 accuracy, except CLIP, achieving 79.3%.

A second experiment was designed to assess the generalization ability of the models to new, previously unseen photos of the same office environment under challenging conditions: flipped and rotated images.

Retrieval performance in this setting is qualitatively analyzed by inspecting whether the true object instance appears among the top retrieved results, as presented in Figure 3.

As we can observe from the Figure 3, in most cases Rank-1 selects the correct object after the embedding comparison. However, when the query object is rotated or flipped, we notice that the performance may degrade and select wrong objects with similar colour. We realize that the high accuracy might be due to the fact of a relatively small dataset, where there is no such a huge variety of objects of the same class. Therefore, to enhance this work, it is suggested to collect a database with more distinct items.

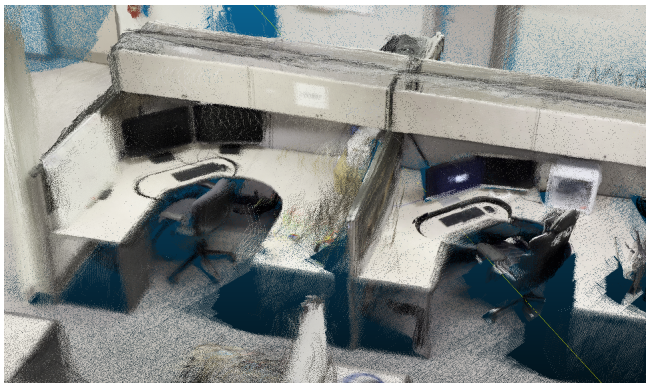

Figure 4. 3D point cloud representation of the office environment, reconstructed by π^3 method.

5. Limitations

This study has several limitations that should be considered when interpreting the results. First, the dataset used in our experiments is relatively small, with a limited number of distinct object instances per class. Although the models are evaluated using retrieval-based metrics such as mAP, Top-1, and Top-5 accuracy, a small number of identities per category can artificially increase performance, since objects of same class are compared only within that class. In larger-scale settings, where many more object identities exist within each category (e.g., hundreds of different keyboards or monitors), the re-identification task becomes substantially more challenging. As a result, we expect that expanding the dataset to include more distinct objects and greater appearance diversity would likely decrease the measured accuracies, providing a more realistic assessment of model generalization.

6. Future Work

From the 521 image frames collected, we also reconstructed a 3D point cloud representation of the office environment (Figure 4) using the π^3 model [13]. This reconstruction provides spatial context that extends beyond the 2D cropped images used for re-identification.

In future work, we plan to integrate 3D data with the Segment Any 3D Gaussian [4] algorithm for semantic Gaussian segmentation. This will allow us to assign semantic labels to 3D regions and precisely locate each identified object within the reconstructed office space. Such a multimodal 2D-3D pipeline would enable one to recognize not only object identities but also their exact physical positions.

7. Conclusion

This study demonstrates the feasibility of applying self-supervised and metric-learning models for object ReID in real-world office environments. By evaluating five representation models on a custom dataset of office objects, we show that feature-based retrieval pipelines can identify object instances using only cropped images and cosine similarity. The results highlight notable performance differences between contrastive, predictive, and triplet-based methods, with DINOv2 and the triplet model showing the highest discrimination quality. Overall, this work provides an initial benchmark for office-object ReID and establishes a foundation for future research on larger datasets, improved detection pipelines, and more robust models.

References

[1] Ali Amiri, Aydin Kaya, and Ali Seydi Keceli. A comprehensive survey on deep-learning-based vehicle re-identification: Models, data sets and challenges, 2024. 1

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. 1, 4

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 1, 4

[4] Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians, 2025. 4

[5] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification, 2017. 1, 4

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. 1

[7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 1, 4

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 4

[9] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 2

[10] Fei Shen, Xiaoyu Du, Liyan Zhang, Xiangbo Shu, and Jinhui Tang. Triplet contrastive representation learning for unsupervised vehicle re-identification, 2023. 1

[11] Fedor Taggenbrock, Ton ten Kate, and Leo Kampmeijer. Robotic surveillance with reidentification for long-term tracking. In *Autonomous Systems for Security and Defence II*, page 136800I. International Society for Optics and Photonics, SPIE, 2025. 1

[12] Changshuo Wang, Xingyu Gao, Meiqing Wu, Siew-Kei Lam, Shuting He, and Prayag Tiwari. Looking clearer with text: A hierarchical context blending network for occluded person re-identification. *IEEE Transactions on Information Forensics and Security*, 20:4296–4307, 2025. 1

[13] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning, 2025. 4

[14] Yihao Wu, Di Zhao, Jingfeng Zhang, and Yun Sing Koh. An individual identity-driven framework for animal re-identification, 2024. 1

[15] Haoxuan Xu, Bo Li, and Guanglin Niu. Identity-aware feature decoupling learning for clothing-change person re-identification, 2025. 1

[16] Bin Yang, Jun Chen, and Mang Ye. Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16870–16879, 2024. 1

[17] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *CoRR*, abs/2001.04193, 2020. 1

[18] Ruiheng Zhang, Zhe Cao, Yan Huang, Shuo Yang, Lixin Xu, and Min Xu. Visible-infrared person re-identification with real-world label noise. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):4857–4869, 2025. 1