Self-Supervised Learning of Pose-Informed Latents

Raphaël Jean[1,2], Pierre-Luc St-Charles[2], Sören Pirk[3], Simon Brodeur[1]

[1]Menya Solutions, [2]Mila, [3]Google Research

Abstract

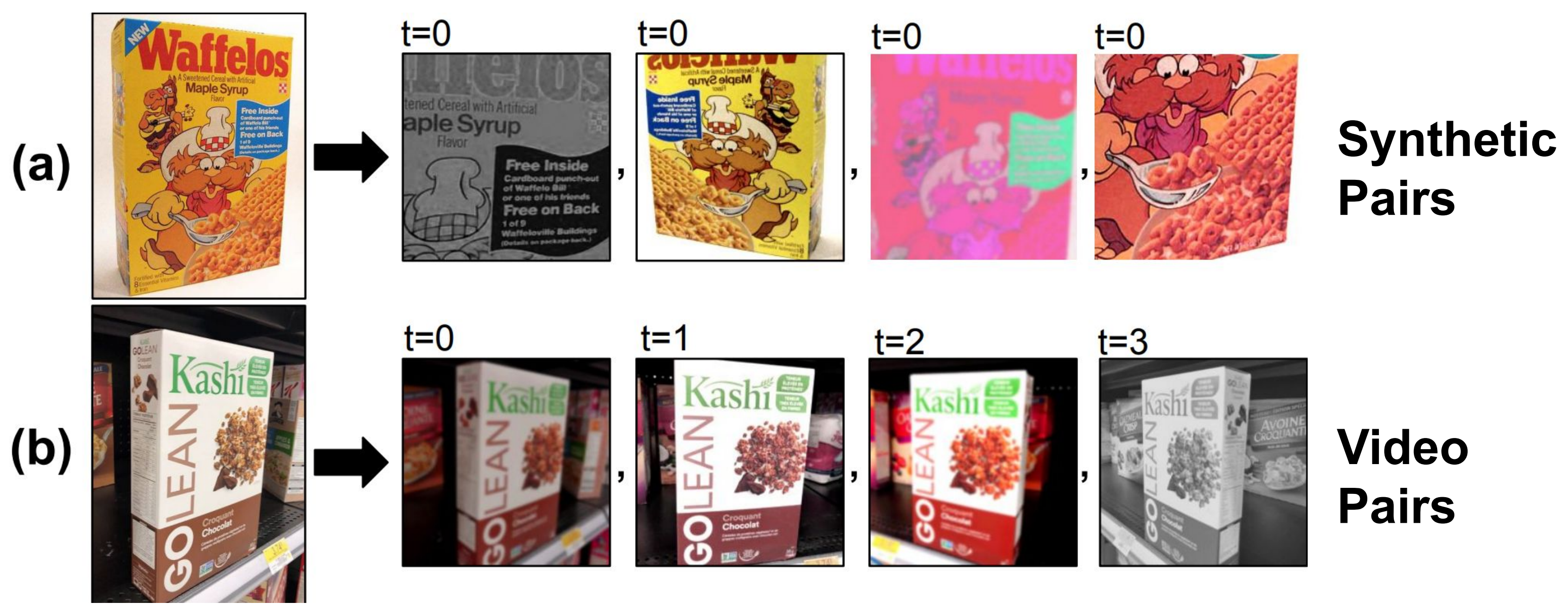
Siamese network architectures trained for self-supervised instance recognition can learn powerful visual representations that are useful in various tasks. Many such approaches maximize the similarity between representations of augmented images of the same object. In this paper, we depart from traditional self-supervised learning benchmarks by defining a novel methodology for new challenging tasks such as zero shot pose estimation. Our goal is to show that common Siamese networks can effectively be trained on frame pairs from video sequences to generate pose-informed representations. Unlike parallel efforts that focus on introducing new image-space operators for data augmentation, we argue that extending the augmentation strategy by using different frames of a video leads to more powerful representations. To show the effectiveness of this approach, we use the Objectron and UCF101 datasets to learn representations and evaluate them on pose estimation, action recognition, and object re-identification. Furthermore, we carefully validate our method against a number of baselines.

Introduction

Background and motivation

Instead of relying **only on synthetic views** for data augmentation when doing **self-supervised learning**, we explore how **videos** can be used to generate news views of the same objects. We study how these new views can affect the quality of the learned representations.

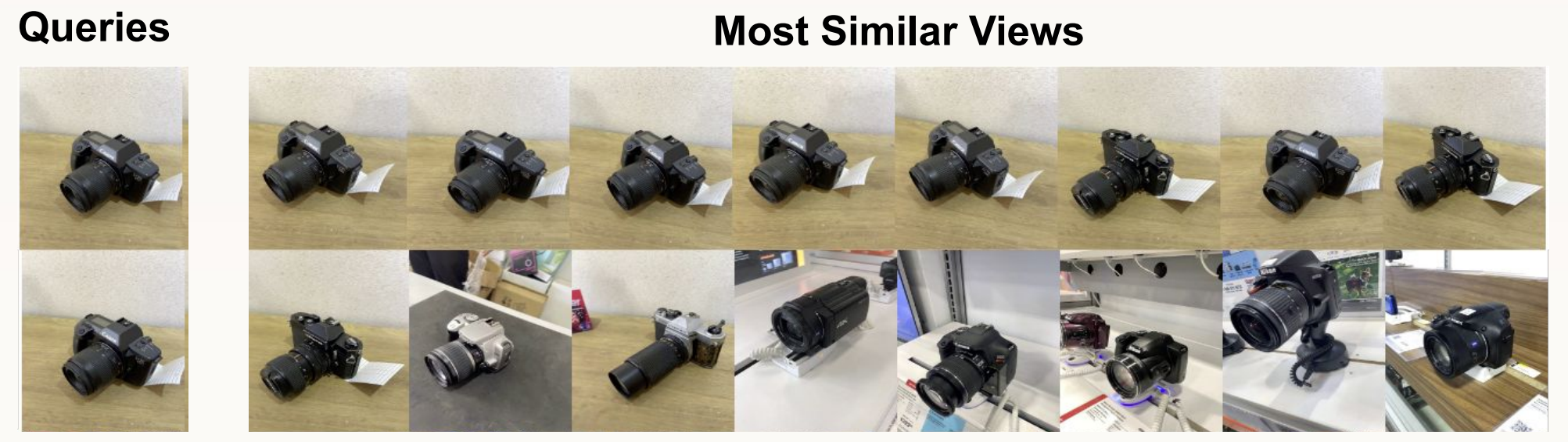
- We **DO NOT** study how videos can be used for video representation learning. This is not our objective. We instead explore **how training with videos** can improve performance in downstream applications that **do not necessarily rely on videos** (e.g. when doing pose estimation from single images).
- Since **Siamese Networks** ([1]) do not require negative view pairs during training, we only focus our study on the analysis of **positive view pairs** generated using pairs of video frames.
- Our results show that positive view pairs created using pairs of video frames can induce either **pose-invariant** or **pose-informed** representations, depending on how temporally distant the frames are.
- Each type of representation is preferable for different tasks; **future works** should try to build an approach that can **extract both representation types at the same time**.



Building Pose-Informed Representations

In order to get pose-informed representation, we pick **pairs** of views with **"similar" pose**, based on the prior that objects views that are **close in time** are also **close in pose**, given a high enough framerate. These views are then used in a standard **SimSiam** training experiment [1].

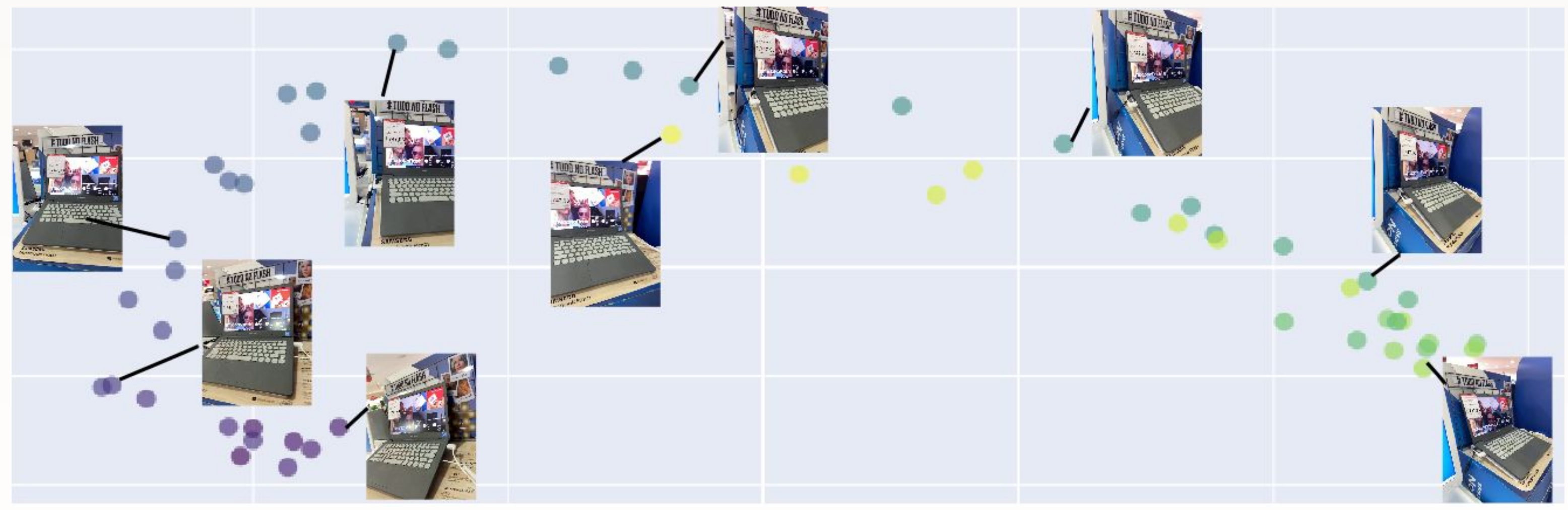
By pulling together views from the same object in **embedding space**, we get representations that are **clustered in terms of object categories, identity, AND pose.**



Results from entire dataset. Neighbors are typically from the same video sequence, with slight pose variations.

Results from other videos only. Neighbors are similar objects in similar settings, or same-category objects, but always with the same pose.

Visualizing object view embeddings using PCA. We observe a "trajectory" in embedding space that seems to correlate with the time axis in the video (shown as the evolving hue of the points). Furthermore, embeddings with similar object poses seem closer in space.



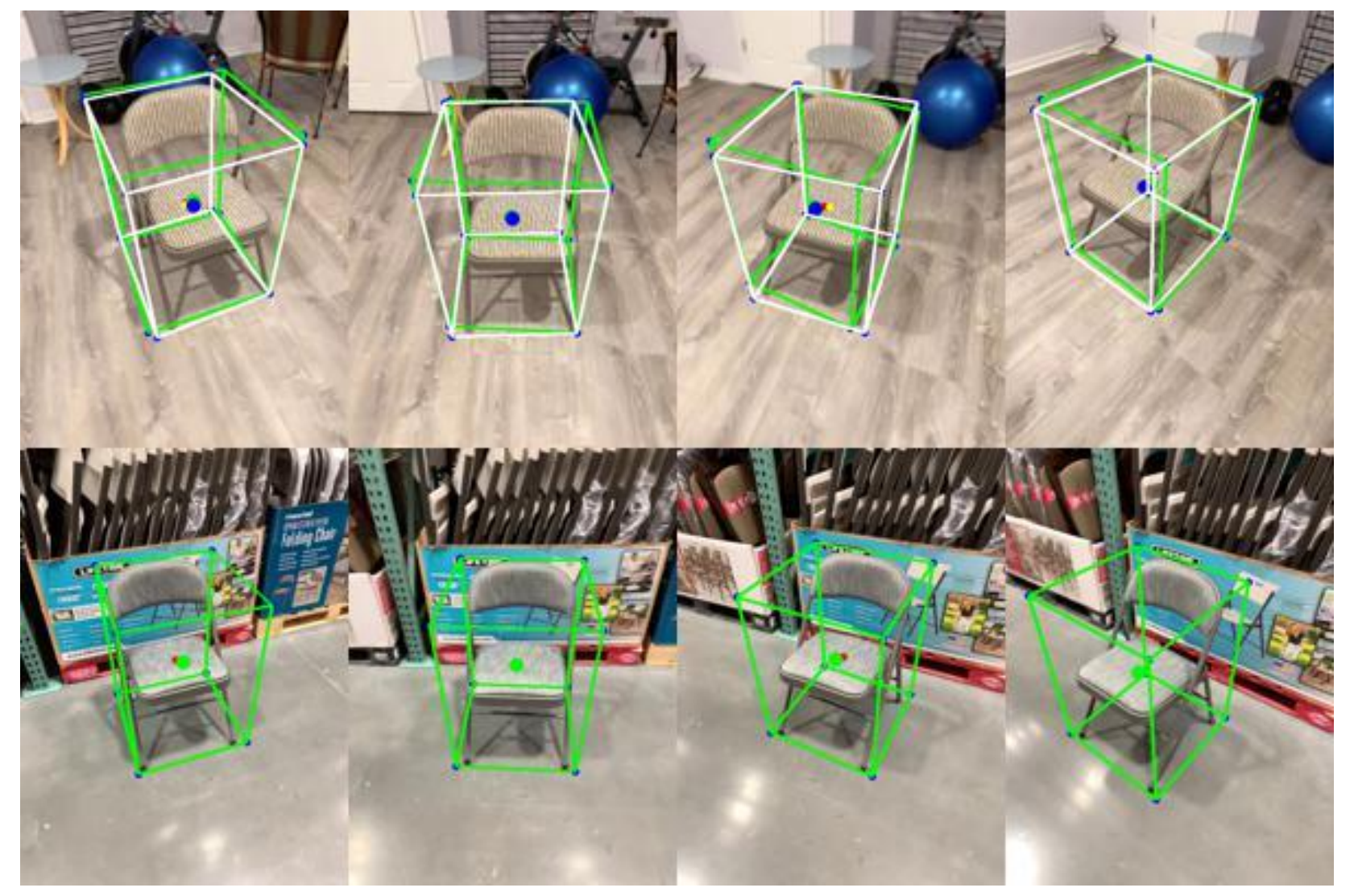
9-DoF Zero-Shot Pose Estimation Benchmark

We tested pose-informed latents on the task of 9-degrees-of-freedom (9-DoF) pose estimation. Our experiments are based on Objectron [2].

This benchmark is **"zero-shot"** since we only get a **single image** of the object whose pose we want to predict **at testing time**. We rely on similar (but not identical) objects found in the training set without using any object category label.

Given an object detector that provides us good **2D object crops** and given a rough **ground plane estimation** in each scene, we seek a **nearest neighbor embedding** in the training set to "fit" a 9-DoF pose to the test frame. Once the embedding is found, we project its corresponding 3D bounding box from the training set to **the ground plane** and scale it using the expected object aspect ratio and scale.

This method allow us to **quantify** how sensitive representations learned with self-supervision are to object geometry and perspective.



In green: groundtruth bounding box. In white: predicted bounding box.

On this line, we show the nearest neighbor that was found for each frame and whose 3D bounding box is fitted in the top scene to generate predictions.

Objectron evaluation results:

Method	bike	book	bottle	cam.	box	chair	cup	lapt.	shoe	overall
Random fit	0.06	0.04	0.04	0.07	0.02	0.10	0.09	0.07	0.05	0.06
Random in-category fit	0.15	0.23	0.26	0.26	0.13	0.46	0.43	0.26	0.11	0.25
Objectron Baseline (1-stage)	0.34	0.18	0.54	0.47	0.55	0.71	0.37	0.55	0.42	0.57
Objectron Baseline (2-stage)	0.61	0.52	0.57	0.80	0.62	**0.85**	0.54	0.67	**0.66**	0.65
ImageNet embeddings	0.46	0.45	0.55	0.74	0.51	0.77	0.72	0.66	0.49	0.59
Our embeddings	**0.65**	**0.54**	**0.60**	**0.82**	**0.68**	0.78	**0.72**	**0.73**	0.66	**0.69**
Our embed. (10% of labels)	0.48	0.46	0.54	0.73	0.48	0.70	0.65	0.66	0.57	0.58
Our embed. (1% of labels)	0.23	0.32	0.47	0.46	0.39	0.61	0.55	0.46	0.30	0.47

Summary:

- Embeddings obtained from an ImageNet pretraining did not provide good matches for bounding box fitting, which shows the usefulness of pose-informed embeddings;
- Fitting with random neighbors did poorly, meaning the bounding box fitting approach itself is not responsible for the overall performance;
- Our zero-shot approach is comparable to fully-supervised approaches in terms of performance (see Objectron 1-stage and 2-stage baselines);
- Our zero-shot approach works even in very limited data regimes (i.e. when fewer training objects are available for the bounding box fitting).

Building Pose-Invariant Representations

Object Re-identification

Assuming we have access to a tracking algorithm and object detector, it becomes possible to extract **multiple views** from the **same objects** with different perspectives. By training on those views, we get models that are more sensitive to the **object (instance) identity** than usual training on static image datasets such as ImageNet.

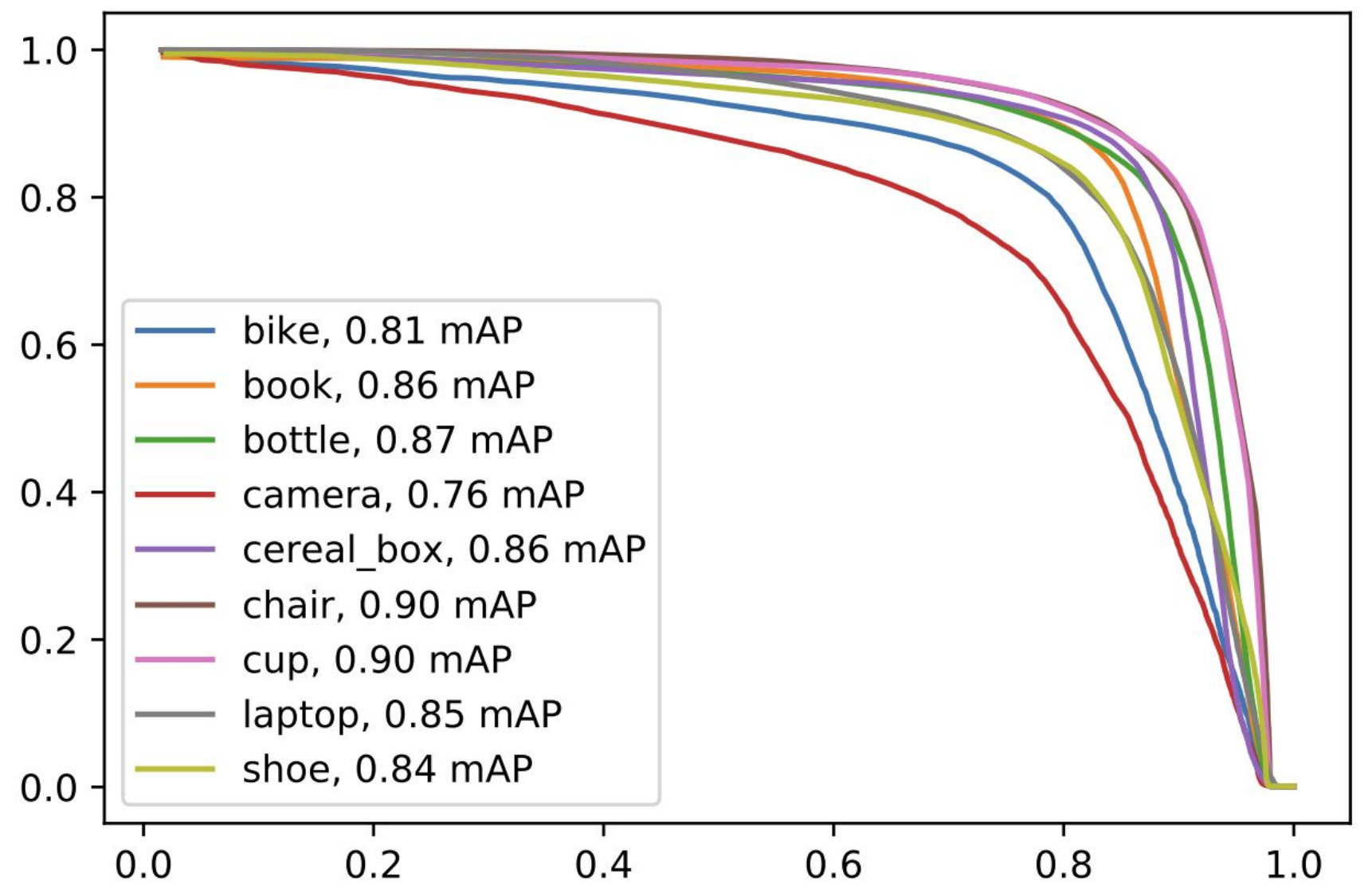


Table 2. Classification, pose estimation, and re-identification results for different setups.

Method	Classif. Acc.	3D mAP	Re-ID mAP
Nearby frame pairs	95.70%	**0.69**	0.53
Distant frame pairs	94.84%	0.58	**0.85**
Same frame pairs	91.19%	0.65	0.26
NT-Xent [8]	**96.36%**	0.64	0.63
TCN [37]	78.20%	0.59	0.19
ImageNet embeddings	96.06%	0.59	0.73
Random embeddings	30.49%	0.17	0.03

Summary:

- Using distant frame pairs as positive views leads to very good representations for instance recognition, which should help object detection and tracking methods;
- Nearby frame pairs lead to lower Re-ID performances, likely due to the lack of variety seen across different views of unique instances;
- The Re-ID performance with distant frame pairs is even better than when using representations from ImageNet that are obtained with orders of magnitudes more unique object instances.

Action Recognition

We assessed the quality of pose-invariant representations on the **UCF-101** action recognition benchmark [3]. This benchmark consist of videos where the **subject** is already **cropped**, and where instances of actions are stored in separate videos. The "action" category for each video is predicted using the nearest neighbor retrieval protocol.

Table 3. Action recognition top-*k* retrieval accuracy on UCF101 for neighbor counts.

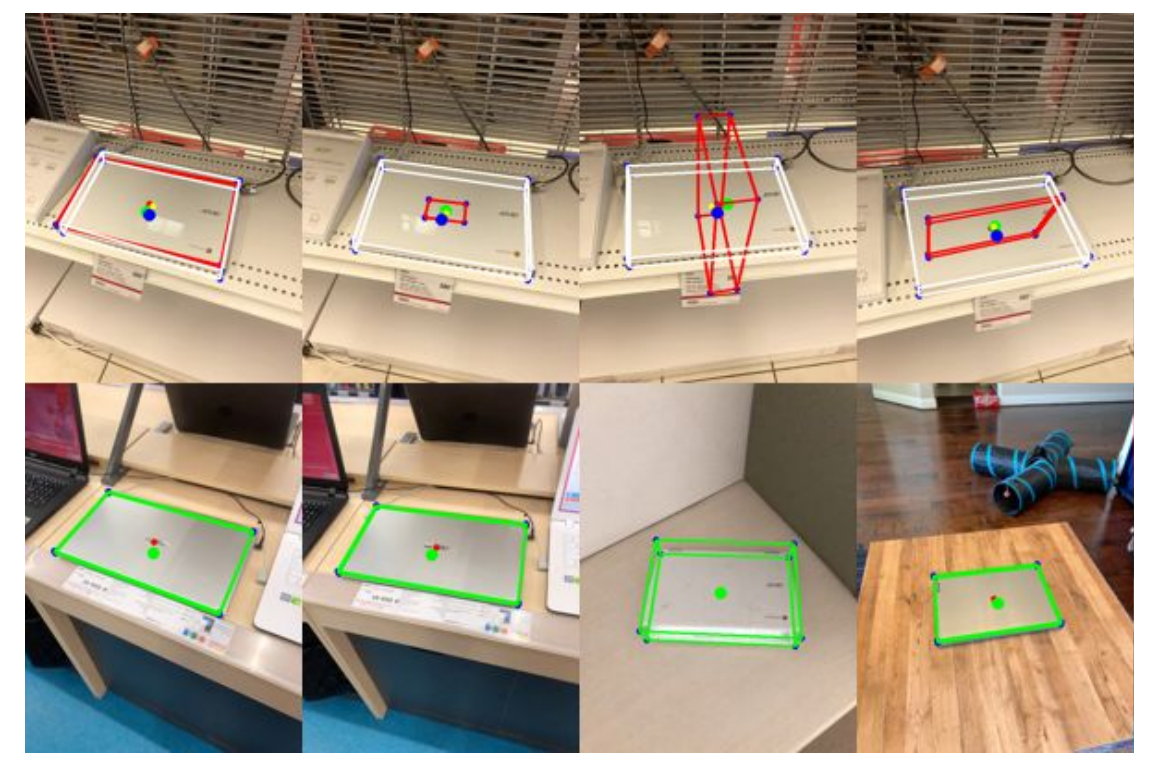
Method	R@1	R@5	R@10	R@20
VCOP [48]	10.7	25.9	35.4	47.3
VCP [30]	19.9	33.7	42.0	50.5
Pace Pred [43]	25.6	42.7	51.3	61.3
CoCLR-2Stream [22]	55.9	70.8	76.9	82.5
Nearby frame pairs	41.4	**57.3**	**62.1**	**66.2**
Single frame	**41.6**	57.1	61.7	65.2

Summary:

- The nearby-frame-pair view generation approach with SimSiam leads to top-*k* recall scores comparable or greater than those of many other self-supervised learning approaches that are specifically made for video representation learning;
- The produced embeddings seem to not only be sensitive to object pose, but also to the motion of objects in a scene.

Conclusion

- Using **frame pairs to create "views"** for self-supervised instance recognition approaches can have **surprising benefits!**
- It would be nice to get the benefits of **both "distant" pairs and "nearby" pairs** in the same representations, or **while training the same model...** (future works!)



Bonus: 9DoF failures! Bad bounding box orientations (far left) or scales (left), and bad ground plane estimations (right and far right) can really break our zero-shot approach.

[1] Chen et al. "Exploring Simple Siamese Representation Learning", arXiv:2011.10566 (2020).
[2] Ahmadyan et al. "Objectron: A large scale dataset of object-centric videos in the wild with pose annotations", CVPR (2021).
[3] Soomro et al. "UCF101: A dataset of 101 human actions classes from videos in the wild", arXiv:1212.0402 (2012).

Code available on GitHub!
github.com/rjean/siampose

This project was supported by a Mitacs grant.