Self-Supervised Learning of Pose-Informed Latents

Raphaël Jean
Menya Solutions, Mila
raphael.jean@rocketmail.com

Pierre-Luc St-Charles
Mila, AMLRT

Sören Pirk
Google Research

Simon Brodeur
Menya Solutions

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.

1. Introduction

The recent progress in unsupervised visual representation learning is owed to multiple orthogonal efforts in improving self-supervision objectives, model architectures, and data transformation techniques. For these objectives, so-called "pretext" tasks were previously used for content understanding through structural manipulations (e.g. rotations [16] and crop-based puzzles [31]). The limited complexity of these tasks resulted in weak feature extractors that would fail to deliver on practical vision tasks. Consequently, they have now been superseded by various forms of instance recognition tasks where each image in a dataset is considered a unique instance. The contrasting of instances across "views" obtained by data augmentation (coined "contrastive learning") has been a very successful approach [7, 8, 32]. Today, various self-supervised learning methods perform on-par with supervised ones on common vision benchmarks. Although it is evident that data transformations play a role in the successes and failures of models, we still do not know exactly what attributes are encoded into scene and object representations as a consequence of their choice. As observed in [3, 8], strong data augmentation that is not helpful in supervised learning is beneficial in self-supervised learning. The augmentation operations that are used essentially dictate which invariances are built into the model's representations. Re-using representations across a wide variety of tasks becomes problematic if the invariances required by the pre-training and downstream objectives are mismatched. A number of methods aim to avoid such issues (e.g. [47]), but these come at an engineering, architectural, or efficiency cost.

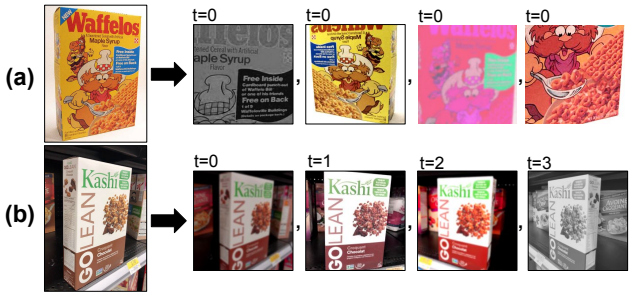

Figure 1. Examples of view generation processes used in self-supervised pre-training: (a) strong data augmentations as proposed in [8] on a single images; (b) simpler augmentations applied to an object viewed across multiple frames. Our experiments show that the latter leads to better representations for pose-sensitive tasks.

In this paper, we advance self-supervised learning by introducing a novel methodology for learning image and scene representations by leveraging the temporal structure of videos. Specifically, we show that a common Siamese network can effectively and efficiently be trained on frame pairs from videos to learn a representation for objects and their surroundings. Depending on the frame pairing strategy used, this representation can be either sensitive or insensitive to the object pose itself, which we refer to as "pose-informed". Please note that we do not aim at video representation learning, but instead only use the temporal structure of videos as a form of weak supervision. This enables us to solve challenging computer vision tasks, such as pose estimation and object re-identification, in a self-supervised

manner.

To this end, we argue that video sequences provide examples of natural geometrical and morphological transformations that cannot be adequately emulated using image-space data augmentation operations (see Figure 1 for a comparison). Our results show that training with diverse views of an object instance enables us to significantly outperform existing methods that only rely on single view training. We also argue that this strategy is more appropriate when combined with recent self-supervised learning approaches that do not rely on negative samples, as the use of these could lead to unwanted biases in the learned representations. We conduct experiments by pre-training on two different video datasets and evaluate the performance of the resulting models on pose estimation, object re-identification, and action recognition tasks. We show that "natural" views obtained from videos are complementary to traditional data augmentation and that they improve model performance across these tasks.

To summarize, our main contributions are: (1) we define a novel methodology for self-supervised training that uses frame pairs from videos as a form of data augmentation strategy with weak supervision; (2) we show that our method can learn both pose-variant and pose-invariant representations (depending on the frame selection strategy used) that can be used to address challenging computer vision tasks for real world applications; (3) our method can directly be used with existing Siamese architectures without the requirement to change the model architecture or loss function; (4) we validate our method on large-scale state-of-the-art datasets, such as Objectron and UCF-101.

2. Related Work

Representation learning is a long-standing research topic in which techniques are proposed to build generic feature extractors for high-dimensional data [5]. These extractors should embed the data in a way that allows high-level attributes (e.g. class labels) to be easily separated or predicted. In practice, they should also be agnostic to the downstream tasks they will tackle, and they should be trainable in an unsupervised fashion. The combination of all these goals is however somewhat paradoxical. In consequence, recent progress has been empirically driven and mostly measured on a small subset of object classification and detection benchmarks. Feature extractors pre-trained on ImageNet rarely perform well when applied to less object- (or instance-) centric tasks such as scene recognition or surface normal estimation [18]. Our goal is to study how view selection could help widen the number of potential applications for pre-trained feature extractors.

Representation learning via instance recognition. As mentioned in the introduction, many researchers have turned their focus towards instance recognition (or discrimination) tasks for self-supervised learning. This kind of task dates back several years [14, 45] but found significant success with the InfoNCE formulation of [32]. Their method uses an autoregressive model to predict embeddings of sequentially structured data across a set of positive and negative samples (or instances). The underlying idea of "contrastive learning" on positive and negative pairs was then substantially simplified in [8, 23]. Recently, alternative methods which are not so dependent on negative samples were proposed [6, 7, 11, 19]. Many of these were also demonstrated to scale well on very large datasets that are less curated than ImageNet [9, 17]. The principles behind the success of all these methods are still the subject of active research [10, 40, 44]. The consensus seems to be that most state-of-the-art unsupervised representation learning approaches rely on the pursuit of two objectives: 1) the representations of different views obtained from a common instance must be aligned; and 2) the set of all representations should be dispersed across the latent space. These two objectives can be explicitly defined as part of the training loss or implicitly through the inductive bias of the model architecture. For our own experiments, we opted to use the SimSiam framework [11] which is a simple formulation that scales well across different training regimes.

On the importance of view generation. Popular representation learning approaches propose different formulations of the instance recognition task, but they all require a view generation process. This process is typically derived from the one of [8]: strong data augmentation is used to make views diverse and thus make the recognition task challenging. In turn, the learned representations become invariant to the set of operations that are used. As discussed in [34, 39, 47], the adopted invariances can sometimes be detrimental to downstream task performance, e.g. rotations and flips may prevent representations from being sensitive to pose variations. According to [39], views that are "good" for downstream performance should only share information that is relevant to the downstream task, which is incompatible with the fundamental goals of general-purpose feature extractors. Furthermore, [4, 15, 28, 29, 33] have noted that independent views obtained by processes that consider the semantic proximity of instances result in much better downstream performance. This motivates our investigation into the usefulness of video sequences for natural and semantically linked view generation.

Representation learning using video sequences. Although numerous works have used video data for representation learning, many of them sought to build feature extractors with 3D CNNs or RNNs specifically for spatiotemporal data analysis [12, 20, 35, 48]. Here, we are more interested in the use of the temporal structure of video sequences as a form of weak supervision for generic representation learning. Early examples of this concept include the work of [42]

which used a representation prediction task as a way to understand and anticipate actions and object detection. The original autoregressive formulation of the contrastive learning approach of [32] also falls into this category: it is a good example of how temporal data can be used to learn good representation for non-temporal tasks. The time-contrastive view sampling strategy of [37] can be easily combined with more recent contrastive learning approaches, but it would still lead to representations that are more time- (and thus action-) sensitive than pose- (or geometry-) sensitive. This is because of the dependency on negative samples that must be extracted from the same video sequence as the positive samples. More recently, [34] investigated temporal view sampling as a way to measure and compare these sensitivity biases on object-centric datasets. For our own study, we investigate the impact of temporal view sampling across a wider range of vision tasks, and using a baseline approach that does not require negative view sampling. We argue that training on video frame pairs leads to the implicit construction of an embedding space that is sensitive to additional object attributes, thus leading to better performance on tasks that also require motion, pose, or geometry awareness.

3. Methodology

In this section, we describe the experimental protocol for the study of a video-based view generation process on self-supervised learning approaches. More specifically, we give an overview of the self-supervised learning framework that we use, the datasets that we pre-train on, the view preparation process itself, and the benchmarks used for evaluation. Note that additional information about the network architecture we used for our experiments, a link to our source code, and additional analysis results are provided as supplementary material.

3.1. Self-Supervised Learning Framework

We rely on SimSiam, the self-supervised learning framework of [11], as a baseline for representation learning. SimSiam distils the idea of view-based instance recognition down to a simple Siamese network architecture where back-propagation is only done on one branch. Many recent self-supervised learning approaches (e.g. [8, 19, 23]) can be retroactively seen as derivations of SimSiam with modifications that increase sample efficiency or that replace its "stop-gradient" with another mechanism to avoid collapses. We choose this particular framework because of its simplicity and because it avoids the need for negative samples.

SimSiam can be described as follows: given an image x and a set of transformations $\mathcal{T}$, we first compute two views $v_1 = t_1(x)$ and $v_2 = t_2(x)$ where $t_1, t_2 \sim \mathcal{T}$. These views are then embedded using a shared backbone encoder (noted f) and projected using a shared MLP head (noted g) so that we obtain feature vectors $z_i = g(f(v_i))$ for $i \in \{1, 2\}$. Next, a prediction MLP head denoted as h is used on one vector to get a prediction of the other. The training loss is defined as the average of the negative cosine similarities between the two pairs of predicted and real vectors, that is:

$$\mathcal{L}(z_1, z_2) = \frac{1}{2}\mathcal{D}(h(z_1), z_2) + \frac{1}{2}\mathcal{D}(h(z_2), z_1), \quad (1)$$

where $\mathcal{D}(a, b)$ measures the negative cosine similarity between a and b. The last important detail is that the gradient of the second term passed to $\mathcal{D}$ is not computed during training, i.e. z_2 is detached from the computational graph when measuring its similarity to $h(z_1)$ (and vice versa for z_1).

3.2. View Preparation Strategies

A key idea of our method is to sample frame pairs from video sequences. The frame pairs should be selected differently depending on what invariance we want to build into the learned representations; we refer to this as "pose-informed". If we choose to sample frames that are close in time, the resulting representations become pose-variant, which is useful for pose estimation. Conversely, if we choose frames randomly from anywhere in the sequence, we get pose invariant representations that are useful for object re-identification or action recognition.

SimSiam, like many approaches before it, relies on a set of data augmentation operations $\mathcal{T}$ to produce multiple views from a single data point. Among typically used augmentations are affine transformations (e.g. rotations, rescaled crops, distortions, flips) that drastically increase the diversity of the views but also result in the suppression of geometric features in the learned representations.

We sample frames in video sequences under the assumption that variations in camera extrinsics create novel views that cannot be equalled with data augmentation alone. On the other hand, if the camera does not move, changes in the appearance of animated objects can still result in views that help understand the behavior of these objects. The combination of both provides an ideal setting for robust representation learning.

With SimSiam, video-based view sampling is fairly straightforward: given an initial frame at time t that we consider as the base image x_1, we select a nearby frame x_2 at time $t + \Delta$, where the time offset Δ is a constant. Then, x_1 and x_2 are augmented into views v_1 and v_2 using a set of "weak" transformations sampled from $\mathcal{T}'$, which is composed of pixel-wise operations (e.g. color jittering, blurring) and distortion-free cropping. Since SimSiam does not require negative views for its objective, we do not have to fetch distant frames from the same video (or random frames from other videos) under the assumption that their semantic link is weaker. This is quite advantageous in reality because it could lead to the suppression of distant temporal relationships between frames which are useful for some tasks

(e.g. action recognition). Using positive views from nearby frames also means that we do not need to track the exact movement of object instances, as the encoder should be robust to partial matches between same-position crops. Our experiments in Section 4.2 assess this robustness by evaluating pre-trained models on re-identification tasks.

3.3. Studied Video Datasets

Representation learning is a data-hungry pursuit, especially when training without supervision. State-of-the-art approaches often rely on the 1M+ images of ImageNet [13], and sometimes even 100 to 1000 times that amount [9, 17]. Interestingly, self-supervised methods can lead to good representations with very little raw data as long as diversified views can be generated [3]. This supports our idea that there may be alternatives to the use of large image datasets. For our experiments, we focus on contrasting the performance of baselines on new tasks and pre-train using video datasets that contain roughly 14,000 clips each. We first use an object-centric dataset that can be tailored into a distribution fairly similar to ImageNet's (Objectron, [2]), and a second not-so-object-centric dataset focused on action recognition (UCF101, [38]).

Objectron. This dataset contains 14,819 videos annotated with 3D bounding boxes over unique objects. These objects are split into nine different categories (bikes, books, bottles, cameras, cereal boxes, chairs, cups, laptops, and shoes) and filmed using mobile phones in an object-centric fashion. The dataset is primarily meant for 3D object detection and localization tasks. The annotations are manually fitted on each object and automatically tracked across videos to ensure every frame has a 3D bounding box. The videos are 10 seconds long on average and mostly feature wide orbital movement around a single focal object. Given proper crops of these objects, the distribution of the frames is fairly similar to the distribution of the corresponding objects in ImageNet. We obtain such crops by simply using the 2D projections of the object's 3D bounding box vertices. In practice, similar results could be obtained by training a category-wise object detector or a salient region proposal network, but we chose to eliminate that source of noise from our experiments to focus on the learned representations. Finally, although the object category sizes are imbalanced, we prefer training on the entire dataset at once instead of training a separate model on each category (as done in [2]). This way, the downstream performance of our model will not be unduly influenced by the knowledge of the imbalanced classes or by easily learnable geometric priors.

UCF101. This dataset contains 13,320 videos annotated with 101 action labels. In contrast with older action recognition datasets, UCF101 proposes a large diversity in terms of video content, recording settings, and action recognition challenges. The action labels are grouped into 5 types, namely: human-object interactions, body-motion only, human-human interaction, playing musical instruments, and sports. Compared to Objectron, it is much less object-centric but much more human-centric, as all actions are related to human activities. The action clips have an average duration of 7 seconds, and we use entire frames as views without cropping.

3.4. Evaluation Benchmarks

Our models are pre-trained without supervision on Objectron or UCF101 and then evaluated for the quality of their learned representations on several "downstream" tasks. These are described next.

Zero-shot pose estimation. We first evaluate the quality of the representations learned on video data by using the Objectron annotations for pose estimation. Given a single crop of a unique (and never-seen-before) object from the test set as a query, we use a pre-trained model to compute an embedding of this crop, search for its nearest neighbor in the training set, and fit the bounding box associated to this result back into the query image. The quality of the representations in terms of pose-awareness is then determined by proxy using 3D bounding box overlap metrics as proposed in [2]. We consider this approach to be zero-shot since the instance and category of the query are not known beforehand. In our case, the pre-trained model is never trained with a pose-related objective, and it remains completely frozen during its evaluation. This is compatible with the standard evaluation procedure used in most self-supervised learning benchmarks [8], and it allows us to compare pre-training strategies without fine-tuning.

Besides, note that the zero-shot formulation of the Objectron benchmark is a 9 degrees-of-freedom (DoF) monocular pose estimation problem. Other popular pose estimation benchmarks (e.g. [25, 26, 46]) focus on the easier 6 DoF problem where instances are the same in the training and test sets. Full pose estimation pipelines for the 9-DoF problem (such as the two proposed in [2]) require the training of a 2D keypoint regression module on top of a backbone. They then rely on object category-aware scale estimates to properly lift the regressed 2D keypoints into 3D bounding boxes. Our zero-shot pose estimation approach allows us to to ignore these extra steps and focus on evaluating the quality of the representations directly. For this, we propose a simple nearest neighbor bounding box fitting algorithm, described next.

Given a crop of a query object and its nearest match in the training set (i.e. the "reference" object), we want to transpose the 3D bounding box of the reference frame back to the query frame. To do so, we use the camera intrinsics matrix and the ground plane estimates provided in Objectron. We first calculate the 3D rotation matrix between the query and reference plane normals. This matrix is used to

roughly place the reference bounding box in the query camera space. Then, we "snap" the bounding box to the query ground plane while aligning the 3D reference object center to its estimated 2D location in the query image. Finally, we repeat the snapping step $n = 3$ times while readjusting the bounding box scale to get a better result. For more information on this algorithm, we refer readers to our implementation which is linked in the supplementary material.

Re-identification and recognition. As a secondary set of benchmark tasks, we evaluate how well the representations learned using video data can be used for the re-identification of instances and the recognition of categories in wide datasets. Such capabilities are especially useful for downstream tasks related to classification and tracking. The pose estimation task does not require that embeddings contain category- or instance-wise features in order to achieve good performance. Simply put, a model that closely embeds all objects with similar shapes, sizes, and orientations would get good results on our pose estimation benchmark, but not necessarily on instance re-identification or category recognition tasks. If our pre-trained models simultaneously perform well on all these tasks, it means that they did not significantly suppress any particular sets of features.

We first assess object re-identification performance using Objectron. We embed all object crops that are part of our validation and test sets and then use cosine similarity to find the nearest neighbors of crops from the test set only. For the evaluation, we compute the Average Precision (AP) using the precision-recall curve for each frame and report the mean; this is referred to as the "Re-ID mAP" in our tables. Moreover, we verify that the object category of the picked nearest neighbors matches the object category of the query object. We report the result using average accuracy across all categories (noted as "Classif. Acc."). Finally, we assess action recognition performance on UCF101 using a nearest neighbor retrieval approach. In this case, we follow the protocol proposed in prior works [12, 21] where the performance is the top-k accuracy for $k = \{1, 5, 10, 20\}$ nearest neighbors. This protocol also avoids having to train a classifier on top of our frozen backbones and instead directly evaluates the usefulness of the frame embeddings.

3.5. Implementation Details

To keep a fairly balanced ratio between the number of frames of the same object and the number of videos of unique objects in both of our datasets, we subsample each video to roughly 5 frames per second. Then, we either set the time offset Δ to its minimum possible value (200 ms), meaning that consecutive frames in the subsampled video can directly form a positive view pair, or uniformly sample frame pairs from the same sequence.

Frame crops are augmented using independent transforms as suggested in [20]. This helps prevent our encoder from simply learning to estimate pixel-level feature flow instead of high-level motion. However, we randomly apply synchronized (video-level) horizontal flips to all views to further increase the diversity of our data. The remaining transformations that are used are a combination of grayscale conversion, color jitter, and gaussian blur, all with parameters similar to the ones proposed in [8].

For Objectron, we use a 90%-10% random split along video sequences for training and validation, and test on the proposed withheld set of videos. As we are using the initial release of the public dataset, we had to preprocess the data to remove instances with invalid annotations (due to TFRecord encoding errors) and blurred input frames (which are a result of the anonymization of personal identifiers in the data). This reduced the original dataset size by roughly 15%.

For SimSiam, our default model configuration relies on a ResNet-50 backbone [24], which is fairly standard across the self-supervised learning literature. The capability of this backbone to scale well with more parameters and larger datasets is already well studied [7–9, 11, 19, 23]. Therefore, we focus our experiments on smaller and faster training regimes that are also better suited to our compute limitations. We rely on the hyperparameter setup of [11] and train using a 10-epoch "warm-up" period followed by cosine annealing over 15 more epochs for Objectron and 90 more epochs for UCF101, which both take roughly the same amount of time (24 hours). Based on preliminary experiments, this training regime results in models that show good baseline performance across all our benchmark tasks.

4. Experiments and Results

In this section we report experimental results that demonstrate the capabilities of pose-informed representations under three different aspects. First, we examine whether our approach leads to representations that encode useful information for pose estimation. Second, we examine whether these representations can still solve other computer vision tasks with good performance. Lastly, we evaluate the impact of using negative samples with our proposed view generation process.

4.1. Qualitative Analysis

We start off with a high-level analysis of our learned representations when using nearby frames as view pairs. In Figure 2, we show how video frames are projected into an intriguingly structured space by a model trained using our proposed methodology. On the left, we can observe that sequences of frames with objects of the same category converge when the object pose becomes more similar. On the right, when focusing on a handful of videos in the same projection, we observe that consecutive frames are co-located and form coherent trajectories. If camera paths diverge, the trajectories bifurcate in the embedding space. This indicates

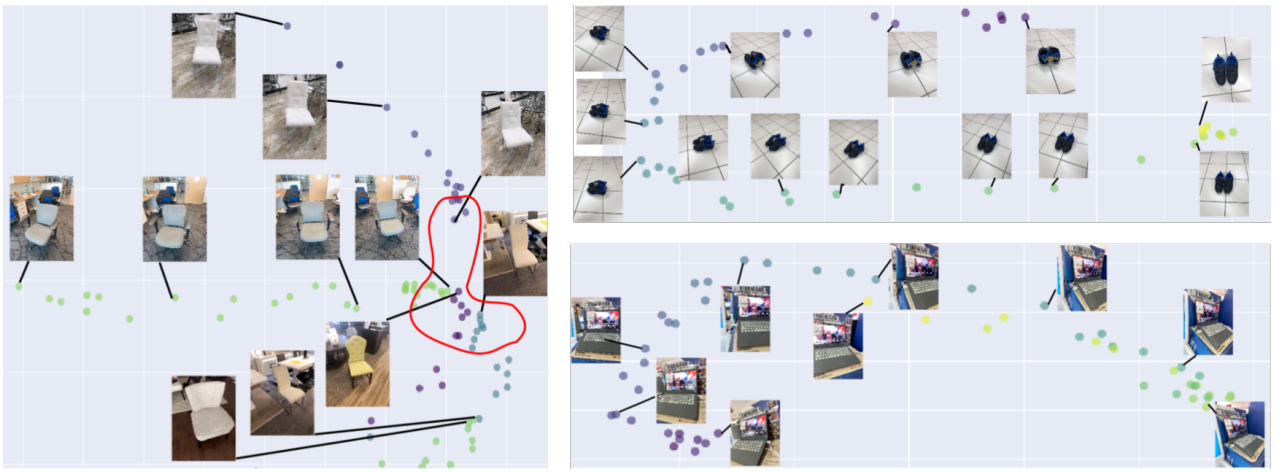

Figure 2. Left: close-up of a t-SNE plot showing the bifurcation of partially co-located trajectories; in the red region, all frames correspond to unique objects (chairs) with roughly the same pose. The frames of separate videos are shown in a different color. Right: two examples of PCA projections of the embeddings of video frame sequences. The color of each point indicates its position in the video sequence (on the viridis scale), which shows the smooth nature of the embedding space.

Figure 3. Nearest neighbors of objects found in the validation set where the leftmost crop in each row shows the query. Top: nearest neighbors including frames from the query video; note that two of the results are outliers as they originate from a different yet similar video (shown with a different border color). Bottom: nearest neighbors across frames of other videos (each with a unique border color); note the diversity in background but the consistency in object poses.

that the trained embeddings can meaningfully encode pose variations.

Next, we visualize the relationship between frames across videos using the learned embedding space. To do so, we embed the entire validation dataset and find the nearest neighbors for individually sampled frames that are used as queries. When a query frame comes from the validation set itself, the nearest neighbors are almost always other frames from the same video (Figure 3, top row). If we omit results from the query video, we observe that the pose and appearance of all the new nearest neighbors are similar to the query, regardless of the background (Figure 3, bottom row). This matching behavior translates very well to the pose estimation task, as bounding boxes of similar shape and orientation can often be found and fitted to queries despite large appearance variations; some examples are shown in Figure 4.

4.2. Quantitative Results

Zero-shot pose estimation. We report in Table 1 the performance of embedding strategies combined with our bounding box matching and alignment algorithm for zero-shot pose estimation. For a comparison with the state-of-the-art, we provide the evaluation results of two recent supervised learning methods: MobilePose [27], denoted as "Objectron Baseline (1-stage)", and an EfficientNet-based keypoint regression approach, denoted as "Objectron Baseline (2-stage)". Both are detailed in [2]. Furthermore, to provide an idea of the lower bound performance on this task with our alignment algorithm, we apply it to randomly selected bounding boxes from the training set ("Random fit"), and to randomly selected bounding boxes from the same object category as the query ("Random in-category fit"). Our results show that our proposed self-supervised training approach outperforms the baselines and state-of-the-art methods in several categories. Given the large differences in model sizes and training regimes with the supervised methods, these results indicate that pose-informed representations provide a significant advantage for pose estimation. We remind readers however that we use "perfect" object detection results derived from bounding box annotations to get our crops, and the performance in practice would be slightly lower when considering object detection noise.

As a quick validation, we evaluated the impact of using smaller sets of usable embeddings: when only 10% of the training video sequences are used for nearest neighbor lookup, the overall performance of our approach is comparable to the *1-stage* baseline. Furthermore, when we reduce the number of used sequences to 1%, we still outperform the *1 stage* baseline on some object categories. Besides, note that the *2-stage* baseline of [2] is actually composed of 9 models in total, i.e. one for each object category. These models were trained in a supervised fashion for a total of 864 GPU-hours on a Tesla V100 [2]. In comparison, our proposed embeddings were obtained after training a sin-

Figure 4. Examples of pose estimation for chairs (left) and shoes (right): our method enables us to obtain similar bounding boxes for pose estimation for each query frame. Furthermore, the fetched nearest neighbors also show similar attributes (e.g. shape) compared to the query frames. The ground truth and obtained 3D bounding boxes are shown as yellow line overlays.

Table 1. 3D mAP @ 0.5 IoU results for zero-shot pose estimation on Objectron. Our results show that our unsupervised learning approach competes with supervised learning baselines [2].

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

gle self-supervised model for only 24 GPU-hours on a RTX 2080 Ti.

Finally, we also report in Table 1 how embeddings obtained using a model pre-trained on ImageNet perform when combined with our bounding box matching and alignment procedure. In short, the performance with these embeddings is decent, i.e. it slightly surpasses the *1-stage* baseline in terms of overall score. However, the *1-stage* baseline is also outperformed in the cup and book categories by the random in-category bounding box fitting approach. This means that our proposed bounding box alignment strategy is quite strong itself. Therefore, the most valuable comparisons we present in Table 1 are between the ImageNet embeddings and our proposed "pose-informed" embeddings, where the latter show a strong upper hand.

Re-identification and recognition on Objectron. Next, we present the evaluation results for classification and re-identification tasks on Objectron in Table 2. We can observe that in the absence of strong data augmentation ("Same frame pairs"), in comparison with a standard self-supervised training methodology (i.e. [8]), we get a marginal improvement in pose estimation accuracy (3D mAP), and a notable decrease in classification and re-identification performance. On the other hand, using pairs

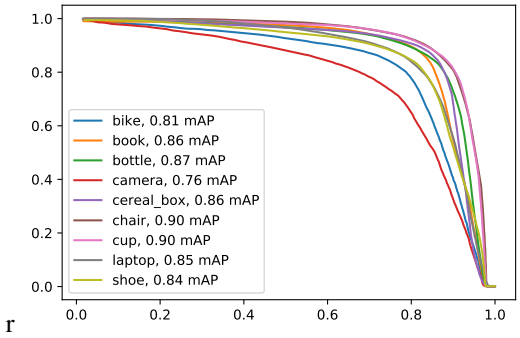

Figure 5. PR curve for our baseline model for object re-identification. Performance is best on categories that have distinctive features across their instances.

across different frames (i.e. "Nearby" which corresponds to sampling with $\Delta = 200$ ms, and "Distant" which corresponds to uniform sampling in the entire video), we get large improvements in either pose or re-identification performance while maintaining high classification accuracy. This shows that using multiple natural views per video sequence is beneficial to train high-performing models. The combination of these results indicates that different-frame-pair embeddings can be used effectively across multiple

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

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

tasks; using nearby frame pairs leads to more pose-aware embeddings, and using distant frame pairs leads to more semantically robust ones. Besides, we note that for object re-identification, the performance is lowest in categories that show less diversity across instances (e.g. cameras); this is illustrated in Figure 5.

Re-identification and recognition on UCF101. We report the action recognition performance on UCF101 based on the nearest neighbor retrieval protocol in Table 3. For comparison purposes, we provide the results of four state-of-the-art self-supervised learning methods [22, 30, 43, 48] that were specifically designed to learn video representations. Here, we can see that simply using non-temporal embeddings learned using SimSiam on individual frames ("Single frame") achieves good performance. Nonetheless, our proposed pose-informed embeddings ("Nearby frame pairs") surpass the classic embeddings when using more than one neighbor. This shows that our embeddings are not only sensitive to camera-object pose relationships, but also to the motion of objects in a scene.

Impact of using negative pairs. To assess the impact of using negative view pairs in the pre-training objective, we evaluated two competing approaches on Objectron; these are shown in Table 2. First, we implemented the time contrastive sampling approach of [37] where negative samples are distant frames from the same video sequence and where the cosine similarity objective of SimSiam is replaced by a triplet loss. This results in worse overall performance than our method, which indicates that sampling frames within a video sequence as negatives results in relevant feature suppression or that the triplet loss is significantly outperformed by the cosine similarity objective. The second approach we investigate is the Normalized Temperature Cross-Entropy ("NT-Xent") loss used in [8]. In this case, negative samples are defined as other view embeddings present in the same minibatch. The results indicate a good average performance across all metrics, but not ideal performance for pose estimation or re-identification compared to using pose-informed embeddings. This shows that some level of feature suppression happens across all tasks when using negative samples.

5. Conclusion

We have introduced a novel methodology for the self-supervised learning of pose-informed representations based on the temporal structure of videos. Specifically, we have shown that a common Siamese network can effectively and efficiently be trained on frame pairs to learn pose-informed representations. We employ a training strategy that maximizes the similarity between views of the same object, which is standard in the self-supervision literature, but we do so by using views of the same object taken at different times in a video.

Pose-informed representations can be used to serve as a foundation for enabling a variety of challenging computer vision tasks where it is important to disentangle camera and object relationships. To showcase the capabilities of our framework, we have trained numerous models on the Objectron and UCF101 datasets and have evaluated them on pose estimation and action recognition tasks. Overall, we observe that using cautious data preparation in conjunction with common Siamese networks allows us to train the networks from scratch more efficiently than previous supervised methods. Consequently, our approach may need fewer training instances due to better data utilization on video datasets.

There are a number of interesting avenues for future works. First, we used what would be considered "large-scale" video datasets for our experiments, but these are still small in comparison to the size and diversity found in large image datasets such as ImageNet. We believe applying the same methodology to much larger video datasets such as YouTube-8M [1] would lead to even more robust and generic representations. Second, our evaluation focused on tasks relevant to the pre-training video datasets, but it would be interesting to evaluate the learned representations on popular benchmarks such as VTAB [49], such as done in [36, 41]. We however expect that our already-trained models would likely underperform on such benchmarks due to the limited diversity of their pre-training datasets, hinting once again that pre-training on larger video datasets is needed. Finally, it would be interesting to confirm whether our training methodology still leads to pose, geometry and motion sensitive representations when applied on datasets of videos that are not as object- or action-centric as Objectron and UCF101.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 8

[2] Adel Ahmadyan, Liangkai Zhang, Jianing Wei, Artsiom Ablavatski, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. *arXiv preprint arXiv:2012.09988*, 2020. 4, 6, 7

[3] Yuki M Asano, Christian Rupprecht, and Andrea Vedaldi. A critical analysis of self-supervision, or what we can learn from a single image. *arXiv preprint arXiv:1904.13132*, 2019. 1, 4

[4] Mehdi Azabou, Mohammad Gheshlaghi Azar, Ran Liu, Chi-Heng Lin, Erik C Johnson, Kiran Bhaskaran-Nair, Max Dabagia, Keith B Hengen, William Gray-Roncal, Michal Valko, et al. Mine your own view: Self-supervised learning through across-sample prediction. *arXiv preprint arXiv:2102.10106*, 2021. 2

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. arxiv e-prints. *arXiv preprint arXiv:1206.5538*, 2012. 2

[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 2

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 1, 2, 5

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2, 3, 4, 5, 7, 8

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 2, 4, 5

[10] Ting Chen and Lala Li. Intriguing properties of contrastive losses. *arXiv preprint arXiv:2011.02803*, 2020. 2

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 2, 3, 5

[12] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *arXiv preprint arXiv:2101.07974*, 2021. 2, 5

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 4

[14] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015. 2

[15] Jonathan Frankle, David J Schwab, Ari S Morcos, et al. Are all negatives created equal in contrastive instance discrimination? *arXiv preprint arXiv:2010.06682*, 2020. 2

[16] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 1

[17] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021. 2, 4

[18] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6391–6400, 2019. 2

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 2, 3, 5

[20] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 5

[21] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. *arXiv preprint arXiv:2008.01065*, 2020. 5

[22] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*, 2020. 8

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 3, 5

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[25] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012. 4

[26] Tomáš Hodaň, Pavel Haluza, Štěpán Obdržálek, Jiří Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017. 4

[27] Tingbo Hou, Adel Ahmadyan, Liangkai Zhang, Jianing Wei, and Matthias Grundmann. Mobilepose: Real-time pose estimation for unseen objects with weak shape supervision, 2020. 6

[28] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. *arXiv preprint arXiv:2011.11765*, 2020. 2

[29] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 2

[30] Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11701–11708, 2020. 8

[31] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 1

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2, 3

[33] Sören Pirk, Mohi Khansari, Yunfei Bai, Corey Lynch, and Pierre Sermanet. Online learning of object representations by appearance space feature alignment. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10473–10479. IEEE, 2020. 2

[34] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *arXiv preprint arXiv:2007.13916*, 2020. 2, 3

[35] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *arXiv preprint arXiv:2008.03800*, 2020. 2

[36] Rob Romijnders, Aravindh Mahendran, Michael Tschannen, Josip Djolonga, Marvin Ritter, Neil Houlsby, and Mario Lucic. Representation learning from videos in-the-wild: An object-centric approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 177–187, 2021. 8

[37] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1134–1141. IEEE, 2018. 3, 8

[38] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4

[39] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020. 2

[40] Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020. 2

[41] Michael Tschannen, Josip Djolonga, Marvin Ritter, Aravindh Mahendran, Neil Houlsby, Sylvain Gelly, and Mario Lucic. Self-supervised learning of video-induced visual invariances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13806–13815, 2020. 8

[42] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–106, 2016. 2

[43] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European Conference on Computer Vision*, pages 504–521. Springer, 2020. 8

[44] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 2

[45] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*, 2018. 2

[46] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 4

[47] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020. 1, 2

[48] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 2, 8

[49] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 8