Self-Supervised Object Detection from Egocentric Videos

Peri Akiva[1,2] Jing Huang[2] Kevin J Liang[2] Rama Kovvuri[2] Xingyu Chen[2]
Matt Feiszli[2] Kristin Dana[1] Tal Hassner[2]
[1]Rutgers University [2]Meta AI

Abstract

Understanding the visual world from human perspectives has been a long-standing challenge in computer vision. Egocentric videos exhibit high scene complexity and irregular motion flows compared to typical video understanding tasks. With the egocentric domain in mind, we address the problem of self-supervised, class-agnostic object detection, aiming to locate all objects in a given view, without ***any*** *annotations or pre-trained weights. Our method, self-supervised object* ***d****etection from* ***e****gocentric* ***vi****deos (****DEVI****), generalizes appearance-based methods to learn features end-to-end that are category-specific and invariant to viewing angle and illumination. Our approach leverages natural human behavior in egocentric perception to sample diverse views of objects for our multi-view and scale-regression losses, and our cluster residual module learns multi-category patches for complex scene understanding. DEVI results in gains up to 4.11%* AP_{50}*, 0.11%* AR_1*, 1.32%* AR_{10}*, and 5.03%* AR_{100} *on recent egocentric datasets, while significantly reducing model complexity. We also demonstrate competitive performance on out-of-domain datasets without additional training or fine-tuning.*

1. Introduction

The ability to detect objects in complex scenes is essential in smart applications and systems, such as autonomous vehicles [31], precision agriculture [3], 3D reconstruction and mapping [58], episodic memory [36], and remote sensing [4]. Broadly stated, the best performing object detection methods require large amounts of densely annotated data, providing bounding boxes for all or most objects in the scene [27, 54, 88]. Such annotations are costly, time consuming to produce, and difficult to scale over large or complex datasets [8]. Recent methods address the costly procedure by using either weak annotations [2, 44, 70, 71], or general self-supervision pre-training [12, 41]. However, such methods lack generalizability to complex scenes, often depending on image-wise features which lack feature granularity, leading to poor object localization and attention coverage. In this work, we aim to both maximize applicable scene complexity and minimize annotation costs by learning a class-agnostic object detector from highly diverse videos without using ***any*** annotations.

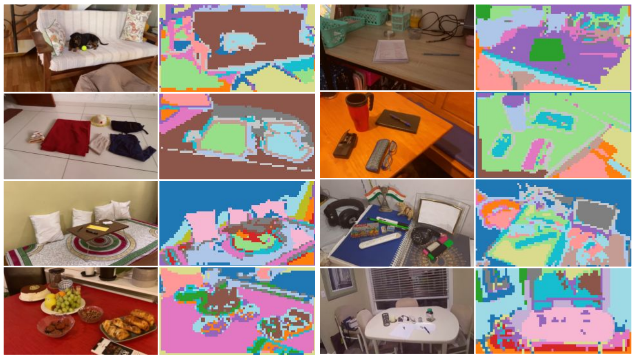

Figure 1. **Image-cluster map pairs.** DEVI learns category-specific, dense features end-to-end from egocentric videos without using ***any*** annotations. Our method is able to distinguish different-category objects, while also remaining consistent for same-category objects. Best viewed in color; colors are random.

We take particular interest in egocentric settings, for several reasons. The first is its complexity: the way humans perceive the world is markedly different from that of many popular datasets (also referred to as "internet images" or exocentric views), resulting in notable new challenges. Internet images [22, 27, 54]–until recently the primary focus of most computer vision methods–capture highly curated, object-centric, specific instances in time removed from global context and filtered from noise and undesired frames; many involve professional and/or manual framing with clear composite objectives. In contrast, egocentric videos typically capture unscripted, "in-the-wild" scenes, replete with dense environments filled with many, diverse objects in varying scales. These significant domain differences result in a weak inductive bias between the internet images domain (*e.g.* COCO [54], ImageNet [22]) and the egocentric domain (*e.g.* Ego4D [36], EpicKitchens [18]), making transfer learning highly difficult [52, 72]; methods designed for and trained on non-egocentric datasets struggle when directly applied to egocentric settings, motivating the development of egocentric-specific methods [9, 19, 51, 67].

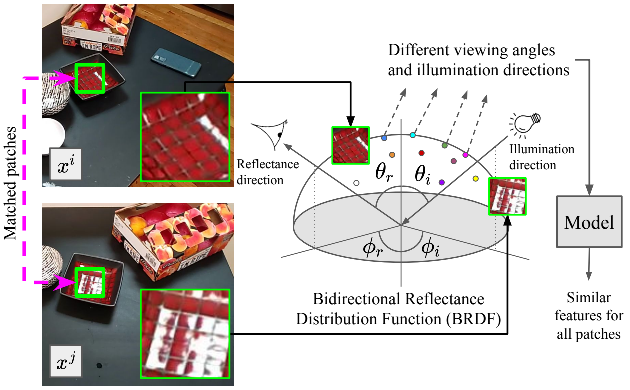
Matched patches
x^i
x^j
Reflectance direction
Different viewing angles and illumination directions
Illumination direction
θ_r
θ_i
ϕ_r
ϕ_i
Bidirectional Reflectance Distribution Function (BRDF)
Model
Similar features for all patches

Figure 2. **BRDF-inspired pipeline.** Given a pair of patches corresponding to the same objects viewed from different viewing angles and illumination conditions, our method seeks to maximize the feature similarity of these patches. Best viewed in color.

We are also interested in egocentric videos for their highly variable viewing directions and egocentric optical flow. While in many ways adding difficulty, these characteristics can provide unique advantages as well, and increased availability of egocentric data in recent years [18, 36, 65, 75] brings new opportunities. In particular, objects and the environment are mostly stationary in many egocentric videos, with head movements and locomotion being the primary sources of camera motion. We draw connections to classical computational appearance-based methods such as the bidirectional reflectance distribution function (BRDF) [62] and Bidirectional Texture Function (BTF) [20], whose computational appearance functions capture the distribution of reflectance measurements of a given opaque surface from all possible viewing angles and illumination conditions at well defined sampling structures and scales. We propose leveraging egocentric camera motion to naturally and unobtrusively sample instances from an object's computational appearance distribution (see Fig. 2).[1] From these diverse views, we propose a novel, end-to-end, self-supervised method for learning features by matching multi-temporal patches covering the same surface or object, thus learning good features for class-agnostic object detection.

An inherent challenge of patch sampling and patch-wise representation learning is content ambiguity: A patch may be sampled from any object, group of objects, or empty surfaces, leading to ambiguous category association. For that reason, we utilize our *object residual module* to encode soft representation to patches, capturing the affinity of patches and all learnt clusters. The object residual module also allows us to define the number of expected categories and therefore learn category-specific features, unlike common self-supervised methods which learn image-wise, general features. To the best of our knowledge, we are the first to learn effective self-supervised features from egocentric videos. We qualitatively demonstrate the ability of our method to generate category-specific features in Fig. 1.

Our contributions in this work are as follows:

1. We present a self-supervised object **d**etection model from **e**gocentric **vi**deos (**DEVI**) that estimates locations of objects in complex scenes.
2. We propose loss functions inspired by computational appearance methods and tuned to egocentric perception named the *multi-view* and *scale-regression* losses.
3. Our object residual module extends existing work on patch representation learning and complex scene understanding to learn category-specific features and precise representation of ambiguous patches without hand-crafted assumptions.

2. Related work

2.1. Self-supervised representation learning

Increasing availability of unlabeled images and videos has inspired researchers to learn effective representations without manual annotations. These unsupervised methods learn invariance to color intensity [47], geometric and affine transformations [61], temporal ordering [28], relative sub-patch localization [24], and patch filling [82]. Traditionally, self-supervised learning (SSL) methods maximize similarity of global features of an image and its transform [13, 14] or learn through clustering features into pseudo-labels [10, 11]. Many of these concepts have been extended to self-supervised video representations as well [30, 37, 38, 39, 64, 68, 76]. Although DEVI and these SSL methods both aim to learn features from unlabeled video, the learning objectives and settings are very different: SSL methods generally learn generic visual features requiring labels to fine-tune on for downstream tasks, while DEVI does not use *any* annotations at any stage of training.

When scenes are complex, with many, diverse, and/or small objects, however, global features fail to capture fine-grain details. To address more complex scenes, recent methods propose patch-wise self-supervision approaches [5, 45, 48] learning local, surface-based feature representations. MATTER [5] does this with remote sensing imagery, though with significant data assumptions, requiring multi-spectral, multi-temporal, and spatially aligned inputs. In contrast to MATTER, we discard the inter-cluster residual weighted average for more concise residual representation, remove the need for spatial alignment by explicitly learning to match patches, and eliminate the multi-spectral constraint by generalizing the notion of material and texture to objects.

[1] More traditional, non-egocentric video datasets such as DAVIS [66] and YT-8K [69] tend to have static or stable cameras following a specific object, and video motion is often due to object motion or actions, with accompanying changes in form or appearance. This makes such video a lesser fit for our BRDF and BTF inspired patching matching method.

2.2. Unsupervised class agnostic object detection

Because of the challenge of converting general self-supervised features to category-specific features, unsupervised object detection is still an open issue. Consequently, many methods [43, 73, 80, 81, 84, 85] follow a naive pipeline similar to the following: (1) self-supervised pre-training of a general purpose network (*e.g*. DINO [12]), (2) generate object discovery predictions for the entire dataset (single object detected, even if there are multiple in the scene), (3) cluster features of discovered objects into a pre-defined number of clusters equal to the number of categories in the dataset (*e.g*. foreground/background for class agnostic object detection), and (4) train an off-the-shelf detector using cluster labels and object discovery predictions.

In order for these methods to be effective, every module in the pipeline must be successful, and individual failures may affect the entire system. While such approaches may be relatively effective on simple datasets such as Pascal VOC [27] or COCO [54], their reliance on object-centric self-supervised pre-training produces sub-optimal features (*e.g*. object saliency) on datasets with increased complexity [36], leading to poor object discovery predictions and subsequent overall object detection performance, as shown in Tab. 1. In contrast, our method trains end-to-end, and is able to learn fine-grain, category-specific features.

2.3. Patch-based learning

Many computer vision methods are based on patch-wise learning, including SIFT [57], HOG [29], convolutional neural networks [50], and vision transformers (ViT) [25]. These approaches use patch-wise features (*e.g*. kernel weights) to obtain global, image-wise representations. Typically, such patch representations are transient or intermediate to some image-level objectives such as detection [86], image-deblurring [60], image-editing [7], or place recognition [40], where global, dense features are achieved. A more explicit utility of patch representations and local descriptors is to predict sparse features, such as keypoints, which are then used for global objectives such as depth estimation [55] and 3D reconstruction [32]. In contrast to these methods, where some patch operation or representation is transient and/or implicit for an image-wise task, we aim to explicitly learn both dense and local features through our *multi-view* and *scale-regression* patch-based loss functions.

2.4. Learning from egocentric data

The unique challenges presented in egocentric data have compelled egocentric-specific methods and data collection efforts for various tasks. The introduction of egocentric datasets such as Ego4D [36] and EpicKitchens [18] propelled work in egocentric action recognition [67], egocentric video-langauge pre-training [53], task understanding [46], and object discovery [9, 19, 51]. Due to the elevated complexity of egocentric settings, methods often require annotated data and/or specialized hardware. Instead, we utilize the innate properties of egocentric videos to implicitly learn high-level object features, without additional annotations, specialized hardware, or intermediate tasks.

3. DEVI

DEVI aims to learn fine-grain, category-specific features that are robust to varying viewing angles and illumination conditions from egocentric videos. We achieve this without ***any*** supervision, pre-trained weights, or hand-crafted assumptions about the data in an end-to-end manner for the task of class agnostic object detection. By using patches, we allow the model to detach local features from their global context and learn patch-level, local objectives, which increase feature granularity and enables isolation of regions in highly complex scenes. Our patch-wise objectives align with our computational appearance analogy: an objective function that operates in the temporal space, enforcing similarity of multi-temporal patches, and a function in the scale space, enforcing similarity of multi-scale patches. The former captures appearance variations in time such as viewing angles and illumination conditions, and the latter captures appearance variations in scale. The framework's training and inference pipelines are illustrated in Fig. 3 and 6.

3.1. Pipeline overview

Given a video $V = \{x^0, x^1, x^2, ..., x^{T-1}\}$ composed of T frames with $x^t \in \mathbb{R}^{3 \times H \times W}$, where t, H, and W represent the time instance, height, and width of the frame. We sample two frames x^τ and $x^{\tau'}$, where $1 \leq \tau' - \tau \leq \delta$, and feed them to two architecturally identical (different weights) transformer-based networks, the patch matching network (Sec. 3.2) and patch feature extractor.

We denote the features of a frame at time t as $\mathbf{z}_s^t \in \mathbb{R}^{L_s \times D}$, representing L_s patches, each a $1 \times D$ vector denoted as $\mathbf{z}_{i,s}^t$ at scale $s \in \mathbf{S}$ and patch location i. For each image patch location i and scale s in time τ, the patch matching network (Sec. 3.2) determines the set of positive indices $P_{i,s}^{\tau \rightarrow \tau'}$ corresponding to the patch matches between x^τ and $x^{\tau'}$, where $|P| \leq L$; we also form a negative set of the negative-matching indices $N_{i,s}^{\tau \rightarrow \tau'}$ (illustrated in the supplementary material). Matched patches in $\mathbf{z}_s^\tau$ and $\mathbf{z}_s^{\tau'}$ are fed to the object residual module (Sec. 3.3) to output residuals $\mathbf{r}_s^\tau$ and $\mathbf{r}_s^{\tau'}$ used for the multi-view and scale-regression losses (Sec. 3.4). We note the sets of anchor, positive, and negative matched patch features as $\mathbf{z}^{\tau +}$, $\mathbf{z}^{\tau' +}$, and $\mathbf{z}^{\tau' -}$, and their residuals as $\mathbf{r}^{\tau +}$, $\mathbf{r}^{\tau' +}$, and $\mathbf{r}^{\tau' -}$, for all scales.

While DEVI requires video input for training, inference can be performed on individual frames (Sec. 3.5). The method assigns multi-scale features at each spatial location to cluster centers learned by the object residual module to

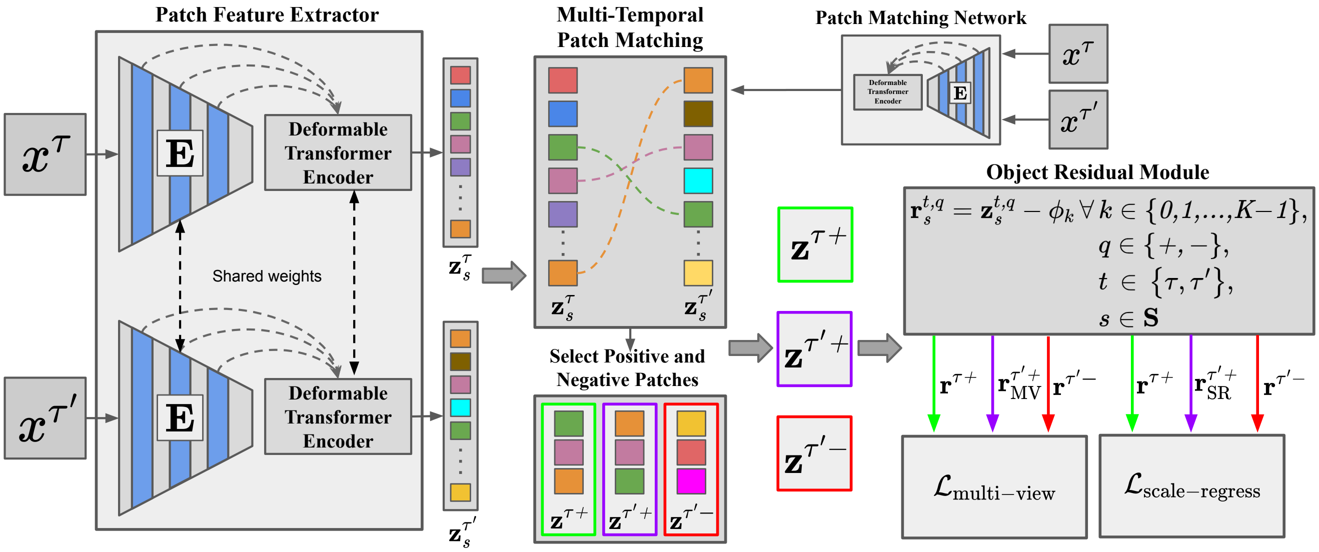
Patch Feature Extractor
x^{τ}
$x^{\tau'}$
E
Deformable Transformer Encoder
Shared weights
$\mathbf{z}_s^{\tau}$
$\mathbf{z}_s^{\tau'}$
Multi-Temporal Patch Matching
Select Positive and Negative Patches
$\mathbf{z}^{\tau+}$
$\mathbf{z}^{\tau'+}$
$\mathbf{z}^{\tau'-}$
Patch Matching Network
Object Residual Module
$\mathbf{r}_s^{t,q} = \mathbf{z}_s^{t,q} - \phi_k \, \forall \, k \in \{0,1,\ldots,K{-}1\},$
$q \in \{+,-\},$
$t \in \{\tau,\tau'\},$
$s \in \mathbf{S}$
$\mathbf{r}^{\tau+}$
$\mathbf{r}_{\text{MV}}^{\tau'+}$
$\mathbf{r}^{\tau'-}$
$\mathbf{r}_{\text{SR}}^{\tau'+}$
$\mathcal{L}_{\text{multi-view}}$
$\mathcal{L}_{\text{scale-regress}}$

Figure 3. **DEVI training framework.** Multi-temporal frames x^{τ} and $x^{\tau'}$ are fed to patch-matching and the patch feature extractor to produce sets of anchor, positive, and negative patch features $\mathbf{z}^{\tau+}$, $\mathbf{z}^{\tau'+}$, and $\mathbf{z}^{\tau'-}$ at all scales $s \in \mathbf{S}$, from which the object residual module generates anchor, positive, and negative residual representations $\mathbf{r}^{\tau+}$, $\mathbf{r}^{\tau'+}$, and $\mathbf{r}^{\tau'-}$, used in the multi-view and scale-regression losses. Positive examples are generated differently for the multi-view and scale-regression losses, as denoted with the MV and SR subscripts, respectively. $\mathbf{E}$, ϕ and K represent the feature extractor, learnt cluster, and total number of clusters respectively. Best viewed in color.

generate spatial cluster maps, from which we extract predicted bounding boxes and confidence scores.

3.2. Matching multi-temporal patches

The task of patch matching is closely related to the task of keypoint matching, which is commonly used for depth estimation [55], 3D reconstruction [32, 78], motion estimation [77], and more. While patch matching is too spatially sparse to be used for these downstream tasks, it provides a few notable advantages for our method: (1) patch matching has additional inter-patch, contextual information compared to a single pixel for keypoint matching, making it significantly easier to learn, and providing more stable predictions under significant view changes, and (2) patch-based architectures and direct patch matching (compared to bootstrapping keypoint matching) provide one-to-one patch correspondence with our patch-based feature extractor. These advantages allow us to train the method in an end-to-end manner by allowing fast convergence for the patch matching task, and simplifying model complexity and overhead operations by re-utilizing the same architecture for both the patch matching and patch feature extractor. The pipeline and integration of the module are illustrated in Fig. 3 and 4.

Training. Given input image x, we apply a random affine transformation, $\mathbb{T}$, on x to obtain $\tilde{x} = \mathbb{T}(x)$. Both x and $\tilde{x}$ are then fed to the patch matching network to obtain patch-wise features $\underline{\mathbf{z}}_s$ and $\tilde{\underline{\mathbf{z}}}_s$ for all scales $s \in \mathbf{S}$. We apply the same affine transformation on $\underline{\mathbf{z}}_s$ to align anchor patches $\mathbb{T}(\underline{\mathbf{z}}_s)$ with their corresponding positive patches $\tilde{\underline{\mathbf{z}}}_s$, while all other, non-corresponding patches are considered negative. Lastly, we use a contrastive loss [63] to enforce feature similarity between anchor and positive patches, and dissimilarity between anchor and negative patches.

Inference. After the training procedure is performed for a small number of epochs, the network weights are frozen, and the model is federated with the main task's training pipeline. Multi-temporal input images x^{τ} and $x^{\tau'}$, are fed to the patch matching network, producing patch-wise $\tilde{\underline{\mathbf{z}}}_s^{\tau}$ and $\tilde{\underline{\mathbf{z}}}_s^{\tau'}$ respectively. We then select anchor, positive, and N negative patch matches based on patch-wise similarity, where anchor and positive patches are the most similar, and anchor and negative patches are least similar (least similar N patches are selected). We note the positive matches indices as $P_{i,s}^{\tau \to \tau'}$ and negative matches indices as $N_{i,s}^{\tau \to \tau'}$. Since the patch matching and patch feature extractor networks are architecturally identical, positive and negative matches indices, $P_{i,s}^{\tau \to \tau'}$ and $N_{i,s}^{\tau \to \tau'}$, have direct correspondence with the patch feature extractor outputs, $\mathbf{z}_s^{\tau}$ and $\mathbf{z}_s^{\tau'}$, for the multi-temporal patch matching.

3.3. Object residual module

An implicit goal of any deep learning method is to cluster features that belong to the same categories closely together. If we consider the task of classification, and scatter image-wise features, it can be observed that examples from a given category are mapped closely together, and are far away from examples belonging to other categories. In order to measure the relative similarity of a given example and all

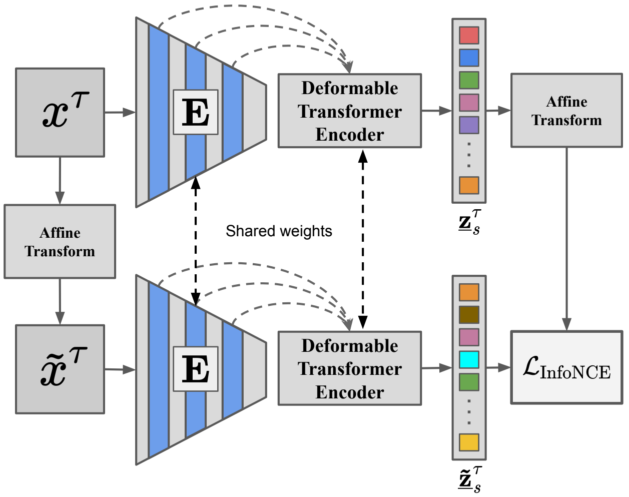
x^τ
E
Deformable Transformer Encoder
Affine Transform
Affine Transform
$\mathbf{z}_s^\tau$
Shared weights
$\tilde{x}^\tau$
E
Deformable Transformer Encoder
$\mathcal{L}_{\text{InfoNCE}}$
$\tilde{\mathbf{z}}_s^\tau$

Figure 4. **Patch matching network training framework.** Image x^τ is fed to a random affine transformation, $\mathbb{T}$, to produce $\tilde{x}^\tau$. Both x^τ and $\tilde{x}^\tau$ are fed to a transformer-based network to produce patch-wise representations $\mathbf{z}_s^\tau$ and $\tilde{\mathbf{z}}_s^\tau$ for all scales $s \in \mathbf{S}$. We then apply $\mathbb{T}$ on $\mathbf{z}_s^\tau$ to produce spatially aligned anchor and positive samples, with all other, non-aligned patches considered negative examples. $\mathbf{E}$ represents the feature extractor. Best viewed in color.

other examples within its own or any other category, we can use residuals. The residual of a given feature vector is the distance or similarity metric from/of that feature vector to a specific cluster center. Recent methods have used residual representation as confidence measurements for classification predictions [33, 42], intermediate soft representation for data quantization [26, 34], and mixed-surface representation learning [5]. Here, we expand upon the work proposed by [5, 42] to learn effective patch representation in highly ambiguous and/or complex environments.

Consider the large-scale patch in $x^{\tau'}_{s=2}$ illustrated in Fig. 5 (in magenta), where multiple objects of different categories are in view (2 bowls and a fruit box). Given our anchor, a small scale patch sampled from $x^{\tau'}_{s=1}$, only depicts one of the bowls, it would be inaccurate to enforce strict equivalence (*e.g.* hard assignment where both patches are labeled 1 and used as ground truth to a cross entropy loss). By utilizing soft representation and residuals, we can represent a patch by its similarity to multiple categories (*i.e.* clusters), which allows us to enforce multi-category similarity and learn from category ambiguous patches.

Given output feature map $\mathbf{z}_s^t \in \mathbb{R}^{L_s \times D}$, and learned cluster centers $\mathbf{\Phi} \in \mathbb{R}^{K \times D}$ with K clusters, each represented by a $1 \times D$ vector. Ideally, clusters centers learn association with specific categories in the dataset, allowing to directly distinguish between objects. The residual of the feature vector z and cluster center ϕ is defined by the distance between them, using $r = z - \phi$. We build a patch-wise residual table, $\mathbf{r}^t \in \mathbb{R}^{L \times K \times D}$, measuring the similarity of all patches in $\mathbf{z}_s^t$ with learned cluster centers using

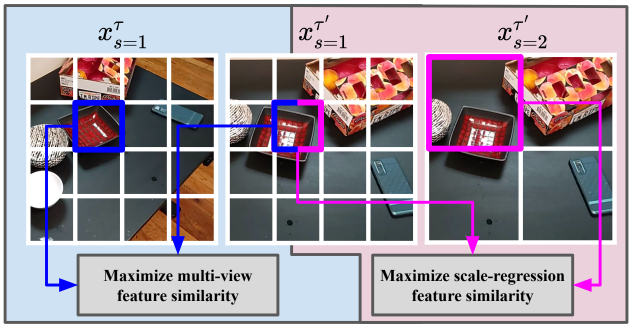
$x^\tau_{s=1}$
$x^{\tau'}_{s=1}$
$x^{\tau'}_{s=2}$
Maximize multi-view feature similarity
Maximize scale-regression feature similarity

Figure 5. **Multi-view and scale-regression losses.** We leverage the natural egocentric perception of human agents to sample diverse perspectives of the same objects. Our multi-view loss maximizes similarity of patch features viewing the same object from different viewing angles, and our scale-regression loss maximizes similarity of multi-scale features of overlapping patches.

$$\mathbf{r}^t = \sigma(\theta||\mathbf{z}_s^t - \phi||_2) * (\mathbf{z}_s^t - \phi) \;\forall\; \phi \in \mathbf{\Phi}, \qquad (1)$$

with learnable parameters $\theta \in \mathbb{R}^{1 \times K}$ and $\mathbf{\Phi}$ corresponding to residual scales and cluster centers respectively, and softmax function σ applied on the normalized residual. We perform this operation on patches in the anchor, positive, and negative patch features sets, $\mathbf{z}^{\tau+}$, $\mathbf{z}^{\tau'+}$, and $\mathbf{z}^{\tau'-}$, to obtain $\mathbf{r}^{\tau+}$, $\mathbf{r}^{\tau'+}$, and $\mathbf{r}^{\tau'-}$, where $\{(\mathbf{r}_i^{\tau+}, \mathbf{r}_i^{\tau'+}) \in \mathbb{R}^{1 \times K \times D}\}$ and $\mathbf{r}_i^{\tau'-} \in \mathbb{R}^{1 \times N \times K \times D}$ for patch at location i, number of negative patches N, and $N < L$. By enforcing similar residual representations of anchor and positive patches, we ensure similar mixture of surfaces and objects within them, increasing robustness to ambiguous scenes.

3.4. Learning similarity across time and scale

Our loss functions aim to leverage the natural egocentric perception of human agents to sample diverse views of the same objects. As humans operate in a given environment, they often either advance towards or circumnavigate objects and elements in their surroundings. The action of circumnavigation allows us to samples multi-temporal patch matches, as described in Sec. 3.2, viewing the same objects from different view points and illumination conditions. These multi-temporal samples are then used for the multi-view loss function, $\mathcal{L}_{\text{multi-view}}$, maximizing similarity of features of corresponding patches. Ideally, this means that objects viewed from different viewing angles, even if visually different (as in Fig. 2), expect to generate highly similar features. We illustrate our loss functions in Fig. 5.

We also address the direct advancement action by proposing the scale-regression loss, $\mathcal{L}_{\text{scale-regress}}$. This loss maximizes feature similarity of a given patch and its overlapping higher-scale patch, increasing model robustness to local viewing scale. How we define positive examples varies between our proposed loss functions. For the multi-view loss, we utilize the multi-temporal patch matching predictions, while for the scale-regression loss, we use the higher-scale patch. Anchor and negative patches remain the same for both losses. Once the sets of anchor, positive, and negative patches are defined, both loss functions (MV=multi-view, SR=scale-regression) are formulated as

$$\mathcal{L}_{\mathrm{MV/SR}} = \underset{0 \leq i \leq |P|}{-\mathbb{E}} \left[\log \frac{\exp(\mathbf{r}_i^{\tau +} \cdot \mathbf{r}_i^{\tau' +})}{\sum_{j=0}^{|P|-1} \exp(\mathbf{r}_i^{\tau +} \cdot \mathbf{r}_j^{\tau' -})} \right]. \tag{2}$$

3.5. Inference from learned clusters

During inference, we first perform per-batch cluster smoothing (batch can be 1 or more frames) on our learned clusters, $\mathbf{\Phi}$, using output features $\mathbf{z}^t$ at time t to obtain smooth cluster centers $\tilde{\mathbf{\Phi}}$. We use the Expectation-Maximization algorithm [21, 59] initialized with $\mathbf{\Phi}$ to iteratively find maximum likelihood cluster assignments for all features in $\mathbf{z}^t$. We optimize this for η iterations (not necessarily until convergence) to produce per-batch smooth cluster centers $\tilde{\mathbf{\Phi}}$ (qualitative examples in supplementary material). This operation allows us to reduce overall noise when assigning features in $\mathbf{z}^t$ to cluster centers $\tilde{\mathbf{\Phi}}$ to obtain cluster map m^t. We separate m^t into a set of blobs using the connected components algorithm [23] and generate bounding boxes around blobs and their confidence scores. The inference pipeline is illustrated in Fig. 6.

Bounding box scoring. Object detection methods traditionally use confidence scores to rank predictions [35]. As we do not use any supervision, usual confidence scores are unavailable: we don't define *what* the method should be confident in. Instead, we define confidence scores of boxes by the convexity of their corresponding blobs, following a common assumption that objects tend to have convex shapes [74]. Given cluster map m^t of frame at time instance t, we use the connected component algorithm [23] on m^t to obtain a set of blobs, b, and their bounding boxes. We define the confidence score of a bounding box, $\mathcal{S}(b_i)$ by the harmonic mean of the convexity measurement and average objectness prior of their corresponding blob b_i, using

$$\mathcal{S}(b_i) = (1+\beta^2) \frac{\frac{\mathrm{Area}(b_i)}{\mathrm{ConvexHull}(b_i)} \mathcal{O}(b_i)}{(\beta^2 \frac{\mathrm{Area}(b_i)}{\mathrm{ConvexHull}(b_i)}) + \mathcal{O}(b_i)}, \tag{3}$$

where β is a scaling factor and $\mathcal{O}(b_i)$ represents the mean objectness prior of blob b_i obtained from an off-the-shelf, self-supervised model [12].

Filtering bounding boxes. We employ classical and unsupervised methods to filter probable false positive predicted boxes, including cluster pruning [16], an objectness prior [49], and convexity thresholding. Cluster pruning is used during the cluster smoothing operation to prune clusters with less than γ mapped pixels. The objectness prior is obtained from an off-the-shelf self-supervised model estimating coarse foreground regions, from which predicted boxes are filtered. Lastly, we employ a convexity threshold ψ, under which bounding boxes are not considered.

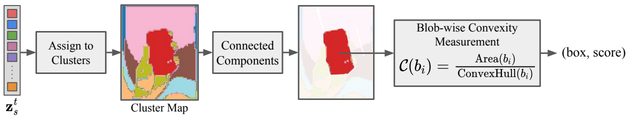
$\mathbf{z}_s^t$
Assign to Clusters
Cluster Map
Connected Components
Blob-wise Convexity Measurement
$\mathcal{C}(b_i) = \frac{\mathrm{Area}(b_i)}{\mathrm{ConvexHull}(b_i)}$
(box, score)

Figure 6. **Inference pipeline.** We assign features to learned cluster centers from all scales to generate cluster map m^t. We then feed m^t to the connected component algorithms to obtain a set of blobs b. We use the convexity measurement function as a scoring mechanism for all $b_i \in b$ and their corresponding bounding boxes.

4. Experiments

4.1. Datasets

We report performance on in-domain (egocentric) and out-of-domain (internet images) datasets of varying complexity. For the egocentric domain, DEVI is trained and evaluated on Ego4D [36] and EgoObjects [65]. Ego4D provides ∼3,600 hours of egocentric videos and is highly complex. For training, we temporally down-sample unannotated Ego4D videos to produce ∼27M frames overall, out of which we use ∼1M. For evaluation, we use the episodic memory validation set, which provides ∼9.9k sparsely annotated frames. EgoObjects [65] provides ∼110 hours of egocentric videos, resulting in ∼66k training frames, and ∼7.7k sparsely annotated validation frames, and is largely object-centric, lower complexity dataset. We emphasize that while we train on video, inference is performed on an image level, making comparisons with other frame-based methods fair. For our out-of-domain (internet images) study (Sec. 5.2), we show our performance on the COCO [54] validation set which provides ∼5k annotated images. Note that we do not train on COCO at any stage.

4.2. Evaluation protocol

We evaluate our method for the task of class agnostic object detection using average precision (AP) and average recall (AR), though we tend to prefer AR (particularly with more proposals) due to the non-exhaustive nature of most object detection datasets [87] making precision measurements less reliable. We use non-maximum suppression with Intersection-over-Union (IoU) threshold of 0.5. Since we use static validation datasets, we run the entire training and evaluation pipeline 5 times and report the mean performance. We note that the difference between

Table 1. **Quantitative results on the egocentric domain.** Average precision (AP) and average recall (AR) on EgoObjects [65] and Ego4D [36] validation sets. DEVI outperforms other self-supervised methods for the task of class agnostic object detection, despite the baselines' increased model complexity and multi-stage procedure. We note the number of stages methods require before final inference.

Dataset		EgoObjects [65]				Ego4D [36]			
Method	# stages	AP$_{50}$	AR$_{1}$	AR$_{10}$	AR$_{100}$	AP$_{50}$	AR$_{1}$	AR$_{10}$	AR$_{100}$
Selective Search [79] IJCV13	1	0.12	0.00	0.86	4.15	0.09	0.00	0.88	3.98
LOST † [73] BMVC21	4	2.13	0.89	6.99	9.41	0.82	0.43	1.42	6.67
FreeSOLO† [84] CVPR22	5	**15.70**	**8.30**	20.70	32.90	2.40	2.80	12.80	17.00
TimeCycle [83] CVPR19	1	6.94	4.24	11.86	12.32	2.95	1.52	6.29	6.97
VideoMAE [76] NeurIPS22	1	9.25	5.16	16.08	16.47	4.05	2.19	8.42	8.96
MoCo V3 [15] ICCV21	1	11.66	7.62	16.83	16.94	4.57	2.59	8.33	8.48
DEVI (Ours)	1	14.96	6.47	**29.61**	**39.43**	**6.51**	**2.91**	**14.12**	**22.03**

multi-object discovery and class-agnostic object detection depends on the evaluated dataset. Multi-object discovery generally refers to when the task is performed on the train set, while class-agnostic object detection refers to when the task is performed on the validation or test set. In this work we only consider the task of class agnostic object detection.

5. Results

Tab. 1 reports the average precision at 0.5 IoU and average recall at 1, 10, and 100 boxes per image for the class agnostic object detection task (additional discussion on metric interpretation is found in the supplemental). We compare with state-of-the-art self-supervised detection methods, LOST [73] and FreeSOLO [84], on the egocentric domain. We train and evaluate these baselines according to reported procedures on Ego4D and EgoObjects, showing the best results. We also compare with recent generic image and video representation learning methods MoCo V3 [15] and VideoMAE [76]. These self-supervised works require fine-tuning on bounding box labels for detection, which we do not assume in our setting; instead, we compare by dropping these pre-trained models in as a replacement to our self-supervised learned patch feature extractor. Our approach outperforms our baselines by up to 4.11% AP_{50}, 0.11% AR_1, 1.32% AR_{10}, and 5.03% AR_{100}, despite their extensive multi-stage and complex pipeline. FreeSOLO requires 3 separate training stages: self-supervised pre-training for object discovery generation (FreeMasks), training on the generated FreeMasks for pseudo-label generation, and then training on the generated pseudo labels for final predictions. This lengthy training process takes $\sim$72 hours of training with substantial computing resources (we use 8 Tesla V100-32GB GPUs), not including any intermediate inference or evaluation steps. In contrast, our method trains end-to-end in $\sim$36 hours with the same computational resources, without any pre-training or multi-training stages, achieving state-of-the-art performance in a single-stage.

Both LOST and FreeSOLO depend on a global, image-wise self-supervised pre-training procedure followed by object discovery (expanded seeded patch for LOST and FreeMask for FreeSOLO). Generic self-supervised methods like MoCo V3 and VideoMAE also tend to have global objectives. When considering dense and complex scenes, as typical in egocentric data, such pre-training strategies result in coarse features, leading to sub-optimal object discovery and class-agnostic object detection. This is supported quantitatively: As scene complexity increases, with COCO and EgoObjects on the lower end of complexity and Ego4D on the higher end, our baselines' performances suffer significantly. In particular, generic self-supervised visual features are not fine-grained enough for detection when bounding box annotations for fine-tuning are unavailable. In contrast, by utilizing patches and residuals, our method has the feature granularity and scene ambiguity robustness to achieve state-of-the-art performance.

In Fig. 7 we present the qualitative results of our method on the EgoObjects and Ego4D datasets. Our method produces bounding boxes that align well with objects in the scene, even when scenes are highly complex. We include implementation details, ablation study, and additional qualitative of results and challenging cases in the supplemental.

5.1. Interpreting metrics

The nature of egocentric data presents challenges not only in network design and increased scene complexity, but also during pre-processing and evaluation steps such as annotations and performance analysis. Due to the high complexity of the data, distinguished by largely varying object scales, diversity, and density, annotation of egocentric videos is often sparse, only considering specific categories at specific scenes. For example, brooms might be annotated when videos are captured in a kitchen, but not annotated when captured in a parking lot. This results in sparsely annotated datasets, which may alter the traditional view on performance metrics. While both recall and precision are affected by the sparse annotation problem, we note that pre-

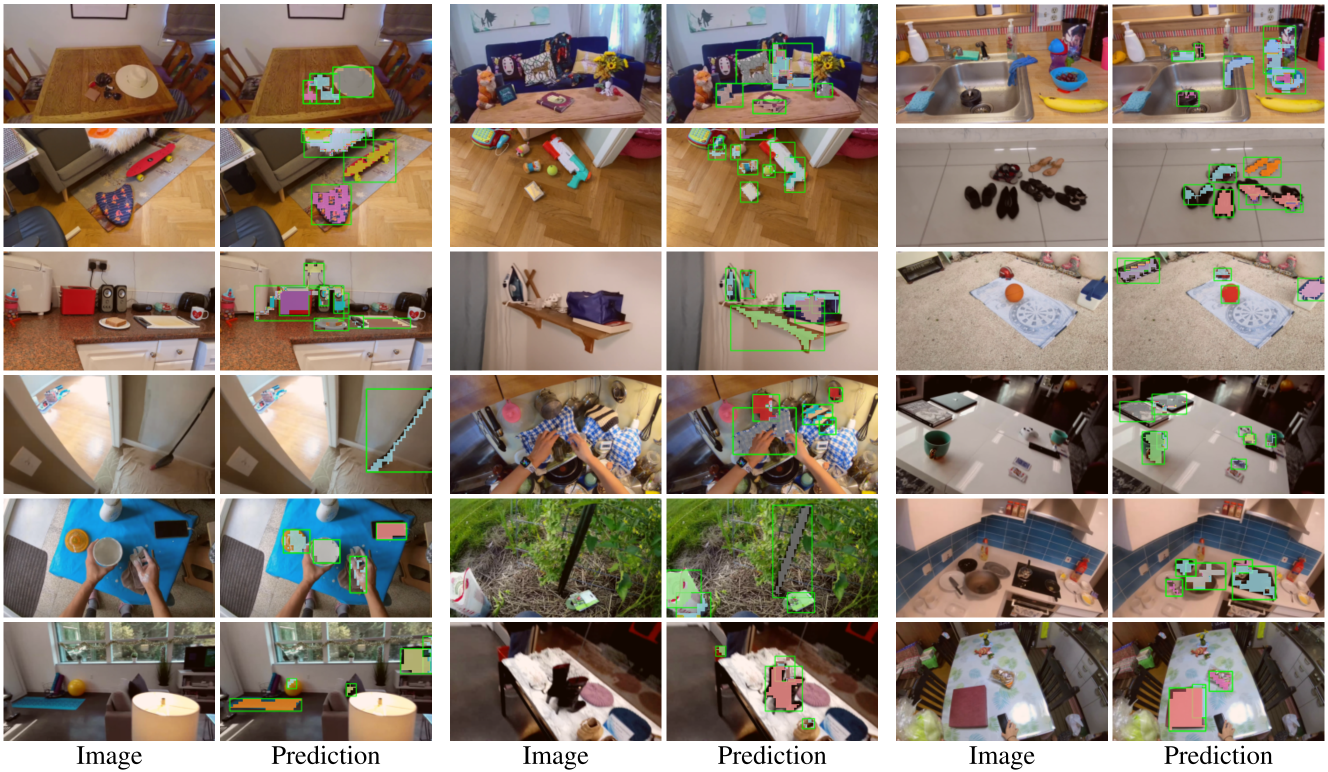

Figure 7. **Qualitative results of DEVI** on EgoObjects (top three rows) and Ego4D (bottom three rows) validation sets. It can be observed that our method has a strong notion of objectness and is able to detect most objects in scenes, even when they are highly complex.

cision is more noticeably affected. That is due to the high number of un-annotated objects in frames, leading to many false positives. As the number of false positives increases, the precision decreases ($P = \frac{TP}{TP+FP}$). For that reason, in this work, we place higher importance on recall performance.

5.2. Ablation studies

DEVI components. To understand what makes DEVI effective, we investigate the impact of various model design choices to overall performance. We study all possible combinations of the loss functions, $\mathcal{L}_{\text{MV}}$ and $\mathcal{L}_{\text{SR}}$, and the Object Residual Module (ORM), and report performance in Table 2. Note that when the ORM is not used, we use K-Means clustering [56] on the raw features instead. Note that at least one loss function is required for training.

We observe worse overall performance without the ORM, which implies its increased utility in ambiguous scenes compared to classical clustering methods such as K-Means. Then, just by incorporating our multi-view loss, $\mathcal{L}_{\text{MV}}$, with the Object Residual Module, we already outperform our baseline by 1.62% AP_{50}. We then further improve our performance by adding the scale-regression loss, $\mathcal{L}_{\text{SR}}$, component. All combinations were trained for the same number of iterations, and with the same hyperparameters.

Table 2. **Ablation study on Ego4D validation set.** We report performance of Object Residual Module (ORM), $\mathcal{L}_{\text{MV}}$ (multi-view loss), and $\mathcal{L}_{\text{SR}}$ (scale-regression loss) model design combinations.

ORM	$\mathcal{L}_{\text{MV}}$	$\mathcal{L}_{\text{SR}}$	AP_{50} (%)	AR_{10} (%)	AR_{100} (%)
		✓	1.12	2.46	8.58
	✓		1.96	4.51	11.87
	✓	✓	2.20	4.05	15.40
✓		✓	2.39	3.31	16.08
✓	✓		4.02	7.92	21.34
✓	✓	✓	**6.51**	**14.12**	**22.03**

Egocentric vs. exocentric training data. While DEVI is designed with egocentric properties in mind, we also investigate the utility of our method on exocentric videos. This aims at verifying our computational appearance-based approach which utilizes the varying viewing angles and illumination conditions of objects in egocentric videos, which may not exist in exocentric videos. Due to the commonly stationary viewing angles captured in exocentric videos, we expect reduced efficacy of output features when trained on exocentric data. For this experiment, we train our model on ∼1M frames from the YouTubeBB-8M video dataset [1] and evaluate on Ego4D validation set. We validate our hypothesis by showing significantly increased performance when trained on egocentric data than on exocentric data,

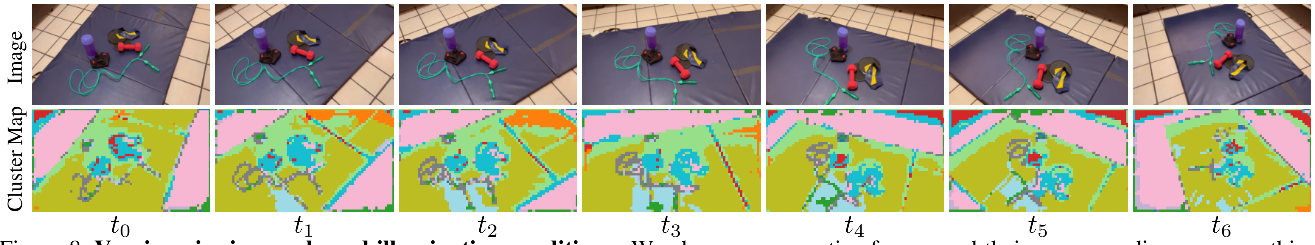

Figure 8. **Varying viewing angle and illumination conditions.** We observe consecutive frames and their corresponding pre-smoothing cluster masks. As the viewing direction changes, we observe that objects retain their cluster assignments, indicating feature consistency regardless of viewing direction and illumination conditions. Colors are random. Best viewed in color.

Table 3. **Out of domain study.** We test our method on unseen, out of domain data to study its generalizability. Note that the baseline methods are trained on COCO while we are not. † corresponds to performance achieved through independent experiments.

Dataset	COCO Validation Set [54]			
Method	AP_{50}	AR_1	AR_{10}	AR_{100}
UP-DETR [17] CVPR21	0.00	0.00	0.00	0.40
Selective Search [79] IJCV13	0.50	0.20	1.50	10.90
DETReg [6] CVPR22	3.10	0.60	3.60	12.70
LOST† [73] BMVC21	4.73	1.99	3.87	8.14
FreeSOLO† [84] CVPR22	9.60	3.70	9.70	12.60
FreeSOLO [84] CVPR22	**12.20**	**4.60**	11.40	15.30
DEVI (Ours)	8.03	3.31	**15.64**	**25.93**

improving AP_{50} by $+3.86\%$, AR_1 by $+1.42\%$, AR_{10} by $+8.80\%$, and AR_{100} by $+16.17\%$.

Out of domain study. We investigate generalizability to out-of-domain datasets by training DEVI on the Ego4D dataset and evaluating on the COCO validation set [54]. We compare DEVI to recent, state-of-the-art class agnostic object detection methods. Note that while our baselines are trained on the COCO training set, our method is *not* exposed to *any* COCO data at any stage. We include reported performance, if available, and performance obtained through our independent experiments (indicated by †). We note that LOST [73] only officially reports performance on the train set, while here we report on the validation set. Despite the domain misalignment, our method is able to achieve competitive performance, outperforming LOST by 3.30%, 1.32%, 11.77%, and 17.70% on the AP_{50}, AR_1, AR_{10}, and AR_{100} metrics respectively (Tab. 3). We also demonstrate competitive performance compared to our independent experiments of FreeSOLO. Qualitative results on the COCO validation set are shown in Fig. 9 and supplementary material.

Varying viewing angles and illumination. We qualitatively visualize our robustness to changes in viewing angle and illumination conditions in Fig. 8; we expect similar cluster assignments for objects across frames. To verify, we visualize the pre-smoothing cluster masks, as it bypasses random elements from smoothing. It can be seen that despite the viewing direction of objects changes (also affecting illumination), the method is still able to retain consistent cluster assignments. Additional examples are provided in the supplementary material.

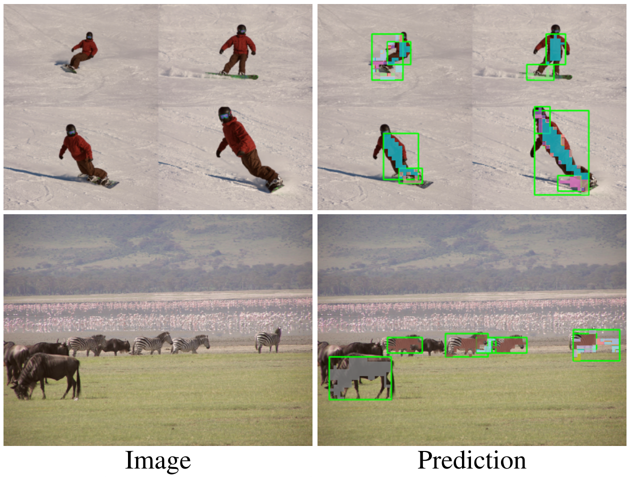

Figure 9. **Qualitative results of DEVI** on COCO validation set. Albeit not trained on COCO, DEVI is able to achieve competitive performance for the task of class agnostic object detection.

6. Conclusion

We have introduced DEVI, a self-supervised class agnostic object detection method for the egocentric domain. We utilize natural human movement patterns to sample views of objects for our computational appearance inspired method, demonstrating that our proposed multi-view and scale-regression losses enable our method to learn robust invariance to viewing angles and illumination conditions. We also show that our object residual module allows learning of effective features in highly complex and ambiguous scenes. Lastly, we achieve state-of-the-art performance on class agnostic object detection on egocentric datasets in a single, end-to-end stage, eliminating lengthy, multi-stage, and computationally expensive procedures.

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] Peri Akiva and Kristin Dana. Single stage weakly supervised semantic segmentation of complex scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5954–5965, January 2023. 1

[3] Peri Akiva, Benjamin Planche, Aditi Roy, Kristin Dana, Peter Oudemans, and Michael Mars. Ai on the bog: Monitoring and evaluating cranberry crop risk. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2493–2502, 2021. 1

[4] Peri Akiva, Matthew Purri, Kristin Dana, Beth Tellman, and Tyler Anderson. H2o-net: Self-supervised flood segmentation via adversarial domain adaptation and label refinement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 111–122, 2021. 1

[5] Peri Akiva, Matthew Purri, and Matthew Leotta. Self-supervised material and texture representation learning for remote sensing tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8203–8215, 2022. 2, 5

[6] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Detreg: Unsupervised pretraining with region priors for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14605–14615, 2022. 9

[7] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 3

[8] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 1

[9] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Unsupervised learning of important objects from first-person videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1956–1964, 2017. 1, 3

[10] 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

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1, 3, 6

[13] 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. 2

[14] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 2

[15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 7

[16] Flavio Chierichetti, Alessandro Panconesi, Prabhakar Raghavan, Mauro Sozio, Alessandro Tiberi, and Eli Upfal. Finding near neighbors through cluster pruning. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 103–112, 2007. 6

[17] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1601–1610, 2021. 9

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3

[19] Dima Damen, Teesid Leelasawassuk, and Walterio Mayol-Cuevas. You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *Computer Vision and Image Understanding*, 149:98–112, 2016. 1, 3

[20] Kristin J Dana, Bram Van Ginneken, Shree K Nayar, and Jan J Koenderink. Reflectance and texture of real-world surfaces. *ACM Transactions On Graphics (TOG)*, 18(1):1–34, 1999. 2

[21] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 6

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[23] Michael B Dillencourt, Hanan Samet, and Markku Tamminen. A general approach to connected-component labeling for arbitrary image representations. *Journal of the ACM (JACM)*, 39(2):253–280, 1992. 6

[24] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2

[25] 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. 3

[26] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021. 5

[27] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 1, 3

[28] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017. 2

[29] William T Freeman and Michal Roth. Orientation histograms for hand gesture recognition. In *International workshop on automatic face and gesture recognition*, volume 12, pages 296–301. Zurich, Switzerland, 1995. 3

[30] Chongjian Ge, Youwei Liang, Yibing Song, Jianbo Jiao, Jue Wang, and Ping Luo. Revitalizing cnn attention via transformers in self-supervised visual representation learning. *Advances in Neural Information Processing Systems*, 34:4193–4206, 2021. 2

[31] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[32] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 963–968. Ieee, 2011. 3, 4

[33] Jan C van Gemert, Jan-Mark Geusebroek, Cor J Veenman, and Arnold WM Smeulders. Kernel codebooks for scene categorization. In *European conference on computer vision*, pages 696–709. Springer, 2008. 5

[34] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Learning representations by predicting bags of visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6928–6938, 2020. 5

[35] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 6

[36] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2, 3, 6, 7

[37] Sheng Guo, Zihua Xiong, Yujie Zhong, Limin Wang, Xiaobo Guo, Bing Han, and Weilin Huang. Cross-architecture self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19270–19279, 2022. 2

[38] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 312–329. Springer, 2020. 2

[39] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. 2

[40] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021. 3

[41] 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. 1

[42] 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

[43] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 123–143. Springer, 2022. 3

[44] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 1

[45] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560, 2020. 2

[46] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35:3343–3360, 2022. 3

[47] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018. 2

[48] Chanyong Jung, Gihyun Kwon, and Jong Chul Ye. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18260–18269, 2022. 2

[49] Tao Kong, Fuchun Sun, Anbang Yao, Huaping Liu, Ming Lu, and Yurong Chen. Ron: Reverse connection with objectness prior networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5936–5944, 2017. 6

[50] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 3

[51] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012. 1, 3

[52] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 1

[53] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 3

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 3, 6, 9

[55] Yebin Liu, Xun Cao, Qionghai Dai, and Wenli Xu. Continuous depth estimation for multi-view stereo. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2121–2128. IEEE, 2009. 3, 4

[56] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 8

[57] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 3

[58] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6851–6860, 2019. 1

[59] Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York, 1988. 6

[60] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *European conference on computer vision*, pages 783–798. Springer, 2014. 3

[61] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2

[62] Fred E Nicodemus. Directional reflectance and emissivity of an opaque surface. *Applied optics*, 4(7):767–775, 1965. 2

[63] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[64] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021. 2

[65] Lorenzo Pellegrini, Chenchen Zhu, Fanyi Xiao, Zhicheng Yan, Antonio Carta, Matthias De Lange, Vincenzo Lomonaco, Roshan Sumbaly, Pau Rodriguez, and David Vazquez. 3rd continual learning workshop challenge on egocentric category and instance level object understanding. *arXiv preprint arXiv:2212.06833*, 2022. 2, 6, 7

[66] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 2

[67] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E2 (go) motion: Motion augmented event stream for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19935–19947, 2022. 1, 3

[68] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 2

[69] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5296–5305, 2017. 2

[70] Ken Sakurada, Mikiya Shibuya, and Weimin Wang. Weakly supervised silhouette-based semantic scene change detection. In *2020 IEEE International conference on robotics and automation (ICRA)*, pages 6861–6867. IEEE, 2020. 1

[71] John B Sigman, Gregory P Spell, Kevin J Liang, and Lawrence Carin. Background adaptive faster r-cnn for semi-supervised convolutional object detection of threats in x-ray images. In *Anomaly Detection and Imaging with X-Rays (ADIX) V*, 2020. 1

[72] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 1

[73] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021. 3, 7, 9

[74] Simon Christoph Stein, Florentin Wörgötter, Markus Schoeler, Jeremie Papon, and Tomas Kulvicius. Convexity based object partitioning for robot applications. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3213–3220. IEEE, 2014. 6

[75] Hao Tang, Kevin J Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. Egotracks: A long-term egocentric visual object tracking dataset. *arXiv preprint arXiv:2301.03213*, 2023. 2

[76] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners

for self-supervised video pre-training. *Advances in neural information processing systems*, 2022. 2, 7

[77] Philip HS Torr and Andrew Zisserman. Feature based methods for structure and motion estimation. In *International workshop on vision algorithms*, pages 278–294. Springer, 1999. 4

[78] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999. 4

[79] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 7, 9

[80] Wouter Van Gansbeke, Simon Vandenhende, and Luc Van Gool. Discovering object masks with transformers for unsupervised semantic segmentation. *arXiv preprint arXiv:2206.06363*, 2022. 3

[81] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *European Conference on Computer Vision*, pages 779–795. Springer, 2020. 3

[82] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11804–11813, 2021. 2

[83] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019. 7

[84] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. Freesolo: Learning to segment objects without annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14176–14186, 2022. 3, 7, 9

[85] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. 3

[86] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021. 3

[87] Yuewei Yang, Kevin J Liang, and Lawrence Carin. Object detection as a positive-unlabeled problem. In *British Machine Vision Conference*, 2020. 6

[88] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1