COHESIV: Contrastive Object and Hand Embeddings for Segmentation In Video

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper we segment hands and the objects they carry by learning contrastive visual embeddings from video. At test time, our system can use a single image and a location of a hand to identify the parts of the scene that the hand is holding. We formulate the learning problem in terms of responsibility, a measure describing how well a hands' motion can explain the rest of the video's optical flow. We train a weakly-supervised neural network using responsibility as a pseudo-label with an attention-based similarity loss and a contrastive loss. Our system outperforms related weakly-supervised methods, achieving good performance on 100DOH and Epic-Kitchen datasets.

1 Introduction

We invite you to pick up an object in your vicinity and bring it towards you. As you hold the object and move your hand, the object moves coherently, and is roughly rigidly attached to a coordinate frame in your hand. While your adult brain does not need this signal and can readily differentiate the object in your hand from both your hand and the background even if you hold your hand still, how did you learn to do this? One answer [40] is that as a human you have time locked modalities of your own hand's configuration via proprioception and vision, and your hand and the object share a "common fate" [46]. The goal of this paper is to operationalize this idea by learning to segment in-hand objects (irrespective of name) from a single image by learning from video data.

This goal poses many challenges for current computer vision. While there has been a long interest in learning segmentation and object individuation from motion cues in vision [39, 32, 11], the general case of broadly being able to segmenting everything has not made substantial progress. While there has, of course, been substantial progress on instance segmentation for particular categories (for instance MaskRCNN [19]) powered by large segmentation datasets [26], the space of categories that one can pick up is vast. Indeed, recent work that aims to reconstruct in-hand objects via segmentation losses [7] reconstructs a few objects by finding category correspondences (e.g., tennis racket masks are used for knives). At the same time, while there is work on understanding hands, including contact state [30] and detecting boxes around held generic objects [38], there is no work on *segmenting* objects, much less work that learns from video data. In sum – segmenting lots of generic objects is still challenging, and segmenting generic objects in contact is no exception.

Our approach makes progress on this challenging problem by assuming a small amount of knowledge about humans. We assume we can estimate landmarks on hands [35], identify whether they are holding something [38], as well as as identify which pixels belong to people [21]. This small amount of information provides a model to guide understanding the rich deluge of information from optical flow [41]. Rather than segmenting the full 3D motions of 3D objects from the flow, we instead only have to identify whether the flow is *better* explained by a moving hand or a background. Moreover,

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

by knowing something about an ubiquitous object (the hand), we show we can learn to segment a large number of far less ubiquitous objects.

We implement our solution with a network, named *COHESIV*, that takes a single image and a point in the image on a hand and segments the object that the hand is holding (Section 3). At test time, *COHESIV* maps the input image to an embedding space with a CNN [36]; this embedding is processed with lightweight heads to produce per-pixel detectors for hands, and feature maps for objects that can be queried to produce a heatmap. At training time, we use the motion with nearby frames to derive signal. We define a *responsibility map* for a moving hand as the relative goodness of fit on optical flow for a planar motion model in comparison to a background model. These responsibilities power two losses: a similarity loss that directly supervises per-hand segments; this loss is supported by a contrastive loss [13] that encourages grouping. Specifically, we use a 3-way intra-image contrastive loss that aims to separate the embeddings of 3 categories of psuedo labels: people, objects, and background.

We train and validate on large-scale video data of humans engaged in complex behaviors (like cooking) using subsets of the 100 Days of Hands (100DOH) [38] and EPIC-Kitchens [9, 10] datasets. We compare our results with a variety of alternate methods, ranging from fully-supervised bounding boxes [38] to basic motion cues from optical flow [41], to saliency [50]. We show that our weakly-supervised method is comparable to the supervised bounding box detector method on all image and hand mIoU, while outperforming RAFT and saliency.

2 Related Work

Our work aims to produce segmentations of in-hand objects from a single image and therefore interacts on a variety of related areas. These range from: the domain we work on (understanding hand-object contact), to the signals we use (common fate with optical flow), to the methods we use to extract meaning from this signal (contrastive learning). Our approach differs from each of these by using contrastive learning to extract supervision from optical flow in order to segment out objects people are holding.

Understanding hands in contact with objects has long been a topic of interest. In addition to considerable work focused on finding hands [3, 29, 38], there has been considerable work on reconstructing hands (for instance with a known shape model like [34]) along with outputs such as poses [12] or 3D meshes [18, 17, 7]. While these approaches can often produce meshes, they usually rely on strong supervision. In contrast, our approach derives its signal from using a small amount of information to make sense of optical flow in videos. While our approach resembles work on human-object relationship detection [44, 16], our target is object in physical contact (e.g., in HOI [44] a person may be using a monitor if they are holding a mouse; in our case, we would only segment the monitor if they were physically grasping it). The most related work is work on understanding objects in contact, which has ranged from highly detailed contact models like [5] to rough bounding boxes like [38]. Our work produces a richer output of a per-pixel segmentation compared to [38] while also requiring less information.

We extract this information from optical flow, which has been recognized as a signal for perceptual organization and recognition since the Gestaltists [46] and Gibson [15]. The utility of optical flow at inference time has been known, for instance for doing motion segmentation ([39, 6, 33, 24, 11, 23] among many others). When coupled with learning machinery, however, the optical flow provides a supervisory signal that does not need to be present at inference time; our work falls into this category. While this signal has been used to learn representations e.g., for pure recognition [32], our approach is most related to work using it for perceptual organization, for instance learning correspondence [45, 22, 47] or boundaries [49, 25]. Our work is most similar to the work learning boundaries but focuses on a specific type of boundary – those of hand-held objects. Rather than learn to segment everything, we learn how to group hands with the objects they hold. To the best of our knowledge, there is no work that does this; the closest is [4], which finds objects that are important for an egocentric user (as opposed to what object is in a person's hand) based on where they look.

Our approach for learning makes use of attention to form predictions and contrastive learning to extract signal from optical flow. We use attention as a means of relating potential query points against an entire image. This type of selective relation is very common in methods that solve visual question answering [1, 28], though our use of attention here aims to instead discover unknown object extents.

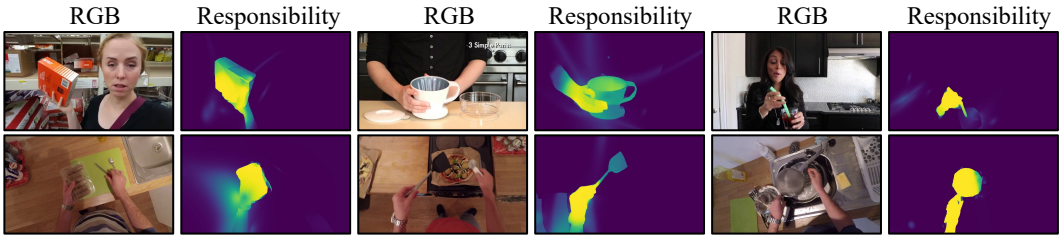

Figure 1: Computed responsibility maps relate individual hands to the extents of their held objects. We illustrate 6 such images and their associated maps, with 3 examples from 100DOH (top) and 3 examples from Epic-Kitchens (bottom).

In this regard, our work is loosely related to [48], an attention-based visual question answering method. Contrastive learning has become a standard formalism of doing self-supervised learning, including remarkable performance by related discriminative methods in separating visual content, [8, 31, 2]. In particular we use a formulation inspired by [42]. While [42] uses saliency to provide a grouping signal on Internet image datasets, our approach uses hands and optical flow. Additionally, our approach, combines self-supervised learning to produce per-pixel embeddings along with a top-down supervision that aims to directly predict what object goes with the hand.

3 Method

Our approach learns to map images and an on-hand query point to predictions of what parts of the scene move with the hand at the query point. While the system requires a single image at test time, it makes use of the rich temporal signal in videos at training time. At the core of our approach is the notion of *responsibilities* for the hands, inspired by the notion of responsibilities when fitting a GMM [14]: the responsibility of a pixel for a hand is how well that hand explains the pixel's motion compared to other hands and the background (Section 3.1).

At test time, we aim to predict this responsibility minus people in new images, which in turn segments out the object a hand is holding. Our network (Section 3.2) predicts this by mapping the image to a joint embedding, which is in turn converted into per-pixel object features that are attended to by per-pixel hand detectors. At training time, the responsibilities provide the basis (along with pixel locations of humans) for a set of losses (Section 3.3) that we use to train our network, using both direct supervision on the outputs as well as a constrastive loss.

Throughout, we assume access to systems that can estimate between-frame optical flow, landmarks on the hand, and which pixels belong to people. We use RAFT for flow [41], projected Frankmocap joints [35] for getting a set of points on hands, and an off-the-shelf person segmentation system [21] for segmenting people.

3.1 Responsibility

We formalize the notion of synchronous motion (or common fate [46]) of hand and objects via the notion of *responsibility*. Given an optical flow $\mathbf{O} \in \mathbb{R}^{H \times W \times 2}$ map and a set of N hands $\mathcal{H}$, we aim to produce N responsibility maps $\mathbf{R} \in \mathbb{R}^{H \times W \times (N+1)}$ with $\sum_{k=1}^{N+1} \mathbf{R}_{i,j,k}$=1 that explain how well each pixel is explained by each hand's motion model or the background. Formally, we the responsibility as a temperature-softened softmax per-pixel, or:

$$\mathbf{R}_{i,j,k} = \frac{\exp_t\left(-d(\mathbf{O}_{i,j}, h_k)\right)}{\exp_t\left(-d_{\text{BG}}(\mathbf{O}_{i,j})\right) + \sum_{k'=1}^{N} \exp_t\left(-d(\mathbf{O}_{i,j}, h_{k'})\right)} \tag{1}$$

where $\exp_t(x) = \exp(-x/t)$ is an exponential with a tunable temperature, d_{BG} and d compute distances between an optical flow vector and a model (namely a background model and a model respectively) and $\mathbf{O}_{i,j} \in \mathbb{R}^2$ is the flow at pixel i, j. Equation 1 requires building and evaluating distances between motion models with respect to a flow vector $\mathbf{o} \in \mathbb{R}^2$. We treat the background as static and thus the model $d_{\text{BR}}(\mathbf{o})$ is simply $||\mathbf{o}||_2$. This is not the best model for egocentric data, but we find it to be effective, potentially due to the high frame rate of the dataset [9] we use.

Our hand models (i.e., $d(\mathbf{o}, h_k)$) assume a *planar* motion because it is simple to solve for, can handle out of image-plane rotation, and is a reasonable approximation at the distances and timescales the data shows. This entails fitting a homography $\mathbf{M}_k \in \mathbb{R}^{3\times 3}$ at landmarks on the hand h_k that should satisfy $[i + \mathbf{o}_1, j + \mathbf{o}_2, 1]^T \equiv \mathbf{M}_k[i, j, 1]^T$ for a point i, j with flow $\mathbf{o}$. We experimented with a number of options, but found the simple approach of fitting $\mathbf{M}_k$ to the estimated flow at Frankmocap [35] joints to be most effective. Using [35] to provide correspondence proved worse, likely since the resulting hand motion model also included disagreement between flow and landmarks as well as between each frame's landmarks. Then, given a model $\mathbf{M}_k$ and pixel i, j with flow $\mathbf{o}$, the models' prediction is $\text{proj}(\mathbf{M}_k[i, j, 1]^T)$ and we define the distance as the difference between the actual flow and the modeled flow,

$$d(\mathbf{O}_{i,j}, h_k) = \left|\left|\mathbf{O}_{i,j} - \text{proj}(\mathbf{M}_k[i, j, 1]^T)\right|\right|, \tag{2}$$

where $\text{proj}(\cdot)$ converts from a homogeneous coordinate to a normal coordinate. While simple, the planar model has a somewhat counter-intuitive catch. By virtue of handling out of plane rotation, the planar model will produce many points that match optical flow by accident specially for static cameras: $\text{model}(i, j) = \text{proj}(\mathbf{M}_k[i, j, 1])$ often has zeros in the image. While a learning system could learn to ignore these accidental matches thanks to training on a large dataset, we aim to accelerate this by preemptively calculating responsibility maps on 6 adjacent frames and averaging.

Techniques: We estimate optical flow using RAFT [41] and estimate hand joints with Frankmocap [35]. Frankmocap fuses a hand detector [38] with a network that predicts 3D hand joints plus weak perspective camera parameters (translation and scale) that allow projection of the 3D hand onto the 2D image.

Data Selection and Implementation Details: We set the responsibility temperature $t = 2$ based on a small held-out set. To obtain averaged responsibility maps, we compute the six responsibility maps using optical flow to the three previous and three next frames.

To ensure that the network primarily sees data where hands are both visible and engaged in contact, we train on clips where: hands appear in 10 consecutive frames; [38] finds reasonably sized and centered hands (between 5% and 50% of the image diagonal width and not within 5% of the image margin); and in contact and moving (hands in contact with objects for ≥ 2 frames and move $\geq 2\%$ of the image diagonal).

3.2 Architecture

The goal of our approach is to take an image $\in \mathbb{R}^{H\times W\times 3}$, a query point (x, y), and produce a segmentation of the object that the hand at (x, y) is in contact with. Our architecture produces a query feature map $\mathbf{Q} \in \mathbb{R}^{H\times W\times A}$ representing hand detectors and a key feature map $\mathbf{K} \in \mathbb{R}^{H\times W\times A}$ representing object features. At test time, one can take a query $\mathbf{q} = \mathbf{Q}_{x,y}$ and produce a detection score at i, j by $\mathbf{q}^T\mathbf{K}_{i,j}$. Thus, a single forward pass can produce the information needed to parse multiple hands.

We produce $\mathbf{Q}$ and $\mathbf{K}$ from a common backbone that produces an embedding $\mathbf{Z} = \mathbb{R}^{H\times W\times F}$. In all but one experiment, we predict the embedding $\mathbf{Z}$ with a standard U-Net-style [36] network with a SE-Net [20] (se-resnext50-4d) backbone with Image-Net [37] pretrained encoder weights. In one cases, we find our contrastive loss to collapse (but to train with a HRNet [43]). Irrespective of backbone, we use two lightweight paths (two 3×3 Conv layers) from $\mathbf{Z}$, concatenated with CoordConv [27], to heads $\mathbf{Q}$ and $\mathbf{K}$. A full description appears in the supplemental.

The use of an intermediate embedding plus the attention heads enables integration of both hand-specific information (e.g., given this hand, these objects go with it) that is handled at the head as well as general information (e.g., these objects tend to not be moved) that is handled at the embedding. The asymmetry via two heads enables producing the *object* the hand is holding rather than the hand: for a single embedding, the hand must have maximum self-similarity if the data is normalized.

Inference: At inference time, the model processes a single image to return embeddings that separate people, objects, and background. When given a query point, it can also return the hand, object, and joint hand-object segmentation masks for that query. This joint inference capability expands the applicability of this model. Beyond segmenting hands and held-objects, it enables an agent to

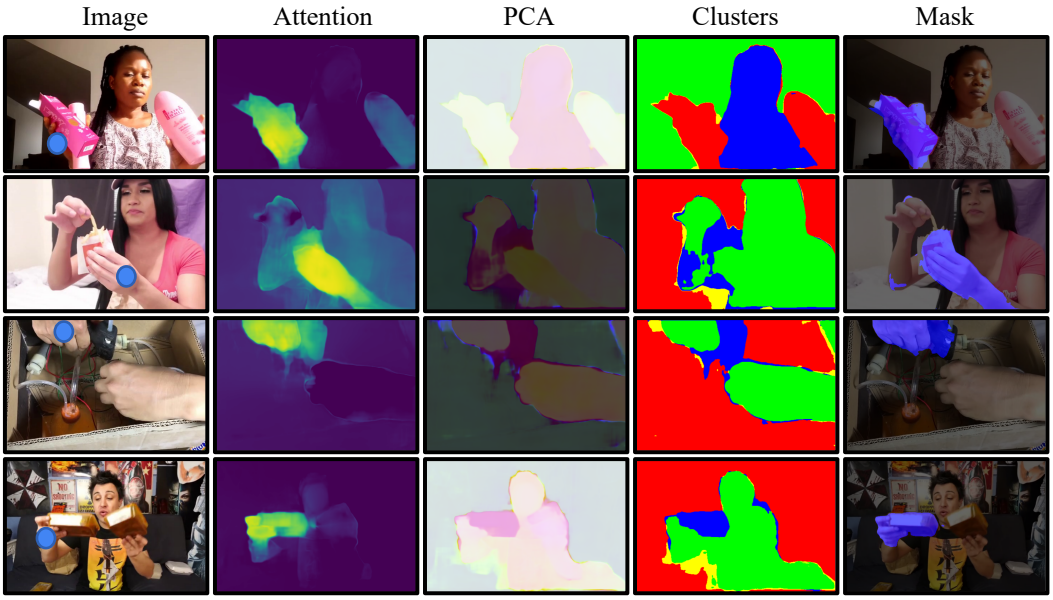

Figure 2: Qualitative results for our method on the 100DOH dataset. We illustrate the input image to the model (column 1), its predicted responsibility map (column 2), a visualization of the 3 first principle components (column 3), a K-means visualization of clustered embeddings (column 4), and the predicted paired hand-object mask (column 5).

query a specific location and understand the extent of that hand's specific held object. We do this by manipulating the resulting embeddings and predicted resposibilities (i.e., $\mathbf{q}^T\mathbf{K}_{i,j}$): we subtract the average detected background and query embeddings. Then, we choose segments of the remaining embeddings that have responsibility above a threshold.

3.3 Losses

We train the architecture with a set of losses. These losses aim to encourage different behaviors for the embeddings and attention heads: the direct query prediction optimizes the loss we care about, while the contrastive loss encourages pixels on hand-held objects to group together. Additionally, these each require dot products between a full embedding space and a handful of vectors as opposed to an expensive construction of a full H^2W^2 tensor comparing all pairwise data. While progress on both losses is often in the same direction, optimally minimizing the proposed losses on the same feature is impossible if there are multiple hands in a scene – the contrastive loss isn't satisfied if all hand-held objects are identical while the direct loss isn't satisfied until each hand-held object is separated. We therefore minimize these losses on *different* features and jointly train.

Attention/Direct Query Prediction Loss: Given a sampled point (x, y) on a hand as well as a responsibility, we can directly supervise on responsibility. Specifically, we compute $\mathbf{Q}_{x,y}^T\mathbf{K}_{i,j}$ for all i, j and supervise. In practice, after we get the prediction from our system, we clip our dot product at 0, putting it in the range [0, 1]. This means that for the system to make something not match a query point, it only needs to make the vectors orthogonal (a large and vast space) as opposed to polar opposites (a very specific location). A per-pixel L1 loss between this predicted value and the computed responsibility maps is then computed, encouraging the predicted results to match the computed responsibility.

Contrastive Loss: We also include a contrastive loss, based loosely on [42], that helps learn more coherent embeddings independent of which hand is being used. These use a set of masks for people (via [21]) and objects (via the union of responsibilities), and background (the remaining pixels). Suppose $\mathbf{z}_P$, $\mathbf{z}_O$, and $\mathbf{z}_B$ represent the average embeddings from $\mathbf{Z}$ for the people, objects, and background masks. Then the contrastive loss aims to ensure that image embeddings within these

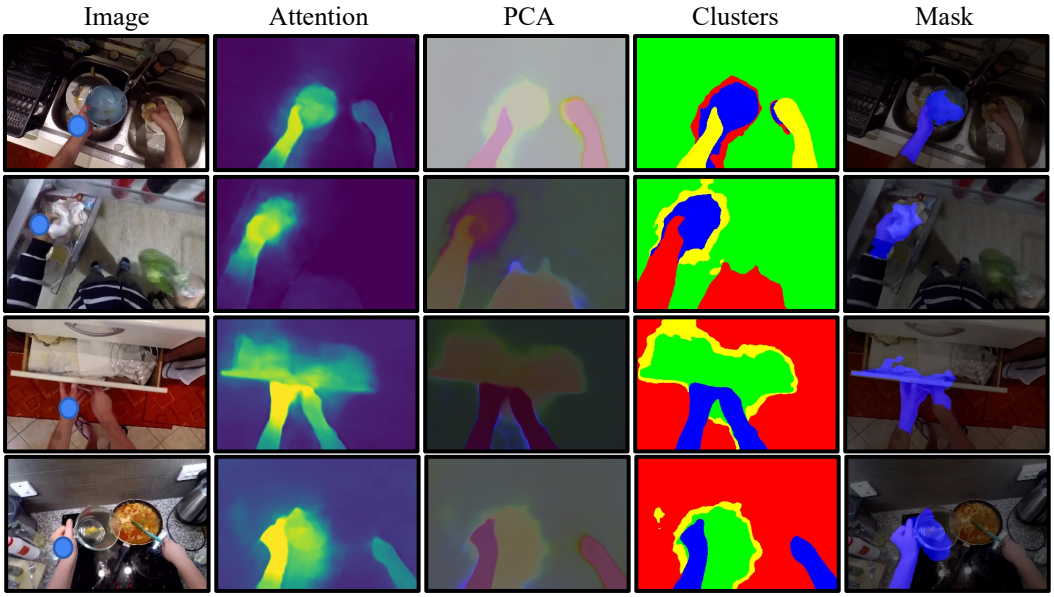

Figure 3: Qualitative results for our method on the Epic-Kitchens dataset. We illustrate the input image to the model (column 1), its predicted responsibility map (column 2), a visualization of the 3 first principle components (column 3), a K-means visualization of clustered embeddings (column 4), and the predicted paired hand-object mask (column 5).

masks are all similar to one another, and that the average embedding across each of these masks is far from the average embedding of the other two groupings. The contrastive loss for hand-held objects at pixel i, j is the negative log-likelihood

$$-\log\left(\frac{\exp\left(\mathbf{z}_O^T \mathbf{Z}_{i,j}\right) + \epsilon}{\exp\left(\mathbf{z}_O^T \mathbf{Z}_{i,j}\right) + \exp\left(\mathbf{z}_P^T \mathbf{Z}_{i,j}\right) + \exp\left(\mathbf{z}_B^T \mathbf{Z}_{i,j}\right) + \epsilon}\right), \tag{3}$$

where $\epsilon = 10^{-3}$ is added for numerical stability. This loss is evaluated only for pixels i, j that are members of the hand-held object pseudolabel. The same loss is defined identically for people and background, simply updating the roles of the numerator query vector and the mask on which the loss is evaluated.

Implementation Details: The masks for people, object, and background are extensions of the responsibility computation described previously. Object mask pseudolabels are responsibility masks that have had the person segmentation subtracted from them, and thresholded at 0.5 to become binary masks. People masks are model outputs from Ternaus [21]. Finally, background masks are the remaining pixels that do not fall into either of the first two categories.

4 Experiments

We evaluate our approach on a data from challenging video datasets, namely 100 Days of Hands [38] and EPIC-Kitchens-55 [9]. The goal of our method is to take an image and a query point on a hand and produce a segmentation for the object that hand is holding. Our primary experiments therefore test how well we can perform that task. We compare with a variety of baselines that test alternate methods of obtaining segmentations for the objects. As we learn to produce the segmentations, we also learn per-pixel embeddings for the scene. We qualitatively evaluate these.

Table 1: Performance on 100DOH and Epic-Kitchens testset using mIoU(%). Our proposed model outperforms all baselines except for the supervised bounding box detector.

	100DOH [38]				EPIC-Kitchens [9]			
	All	Pair	Hand	Object	All	Pair	Hand	Object
COHESIV	47.8	42.0	50.4	15.5	34.4*	29.1*	54.7*	17.9*
Bbox [38]	56.9	47.0	56.5	34.9	54.3	44.8	53.8	34.4
Saliency [50]	23.4	-	-	-	17.4	-	-	-
Flow [41]	18.5	-	-	-	19.5	-	-	-

4.1 Datasets and Experimental Setup

Datasets: We first generate 80.4K 10-frame training clips from a subsection of 100DOH [38] dataset. This dataset is from a large variety of public YouTube videos, that were uploaded by users and then collected into a dataset and contains people, but has been screened for offensive content. Specifically, we select from the "make drinks" and "make food" videos due to the variety of objects and the tendency of these objects to be roughly rigid We also generate 30K 10-frame training clips from action segments in Epic-Kitchens [9] dataset. This datasets was collected in a lab and only shows peoples' hands (but also their home environments) and does not have offensive content. Given each 10-frame clip, we generate and filter responsibility maps for two middle ($5th$ and $6th$) frames for our training tasks. Full details are in the supplemental.

New Contact Segmentation Annotations: To evaluate our approach, we gather a new set of labels through a 2-step manual labeling, performed by an annotation company. Given a red dot on each hand, we ask annotators to label the hand mask; then, given the hand mask, we ask annotors to label the object mask if it exists. This yields 1,123 test images (2,105 hands and 1,810 objects) and 482 validation images (889 hands and 775 objects) for 100DOH [38] and 1,170 test images (2,033 hands and 1,791 objects) and 438 validation images (735 hands and 639 objects) for Epic-Kitchens [9] This data is sufficiently large to validate our approach (although not enough for training).

Training details: We trained a *COHESIV* model and an Attention-only and Embeddings-only ablation each on 5 GTX 1080 GPUs. We use the AdamW optimizer, with initial learning rate between 1×10^{-2} and 2×10^{-2} and batch sizes of 2 images per GPU. We used a learning rate scheduler which cut the learning rate in half when validation loss failed to decrease across 2 epochs. We also used early stopping, ending the experiment if validation loss failed to drop after 3 epochs. We found that our contrastive loss to collapse when trained with the attention on with the [20] backbone, on EPIC-Kitchens and so substitute a similar HRNet [43].

Metrics: Our goal is to take a single input image and a query point (x, y) corresponding to a hand and produce a segmentation. We use intersection over union as our base evaluation metric (and use grid search on validation set to set thresholds for binarization). Since different parts of our approach (e.g., the embeddings vs. attention) produce different types of outputs, we evaluate this in four settings, done by changing the definition of ground-truth positive: **All:** any hand or object is considered a positive, which evaluates how well a method can identify hands and hand-held objects; **Pair:** the hand in question and the object (if any) it holds are considered positives, which tests whether a system can individuate hands; **Hand:** the hand in question is considered a positive; **Object:** the object the hand is holding (if any) is considered a positive. While COHESIV can solve all four challenges through combinations of its outputs, different methods may or may not be able to. When it is impossible, we denote this with –.

4.2 Contact Segmentation

We begin by evaluating contact segmentation. Our system produces a rich space of outputs through its embedding as well as query and key responsibility prediction system. Our inference procedure (under *Inference* in Section 3.2) uses these embeddings to predict hands and held-objects both for a specific query point and for all points.

Table 2: Ablation performance on embeddings, attention, and combined embeddings and attention on 100DOH and Epic-Kitchens using mIoU(%). It is worth noting that the solely attention based model cannot readily separate the co-occuring motion of hand and objects

	100DOH [38]				EPIC-Kitchens [9]			
100DOH	All	Pair	Hand	Object	All	Pair	Hand	Object
COHESIV	47.8	42.0	50.4	15.5	34.4*	29.1*	54.7*	17.9*
Attention Only	40.1	37.2	-	-	39.2	38.7	-	-
Embeddings Only	20.5	16.1	13.0	10.6	28.8	20.0	17.7	14.7

Qualitative Results: Figure 3 shows the full output of the COHESIV system. Column 2 (Attention) shows responsibility predictions (vidiris) for the given query location (blue dot in input). Qualitatively, our results are effective at capturing the extents of held objects. For instance, row 1 depicts an image with two hands and two held objects. Our system successfully embeds both objects as held objects, and then our attention prediction successfully localizes the hand and held object under the query dot. Row 2 shows a complicated segmentation where objects held by different hands are overlapping, yet our system successfully identifies the boundaries of the correct held object. At times, large objects are poorly predicted. In such instances, the object extents shown in column 3 and 4 (PCA and KMeans respectively) serve to improve the responsibility result, capably segmenting objects.

Baselines and Ablations: We include 3 baselines to understand the capabilities of the system. (*BBox*): The first baseline is the supervised hand-object bounding box detector from [38]. Since this is trained on over 132K annotated examples from the training sets of both 100DOH [38] and a variety of egocentric datasets, including [9], we do not expect to match its performance. We report it, however, for context. Of course, our system by virtue of being segmentation, can in principle get improved segmentations (although the box is often a good approximation for the object). (*Saliency*): Our second baseline is a recent saliency detection system [50] which aims to identify regions of interest in an image, and serves as a benchmark for how foreground/background segementation approaches might perform on our task. (*Flow*): Our final baseline is an optical flow baseline. We use RAFT [41] to estimate the flow between adjacent pairs of video in an image, and then treat all flow above the mean value as a foreground prediction. Finally, we report ablations of using just our attention loss and just our embedding loss. The saliency and optical flow baselines aim to serve as alternate explanations for the results seen (e.g., that the approach is just picking up on a salient object).

Quantitative Results: In Table 1, we present how our model compares to similar baseline methods. Our approach approaches the performance of the supervised approach of [38], and our systems outperform more simple approaches of finding salient objects or looking for above average flow. Our COHESIV model on EPIC-Kitchens is trained with a HRNet [43] backbone, since we found the SE-Net [20] backbone to train poorly with both losses on EPIC-Kitchens. We note, however, that the HRNet performs similarly on 100DOH (within a few percentage points).

In Table 2, we demonstrate the value of our hybrid attention+embeddings model, by comparing against ablated versions. Jointly training generally improves results on 100DOH. On EPIC-Kitchens, joint training slightly degrades results compared to the attention only system, although this may also be explained by the change in backbone. Nonetheless, our prediction of both embeddings and attention-style detection enables the system to produce full outputs.

4.3 Analysis and Limitations

While we cannot directly evaluate the learned embeddings, we show results throughout the paper using both PCA and clustering. For PCA, we fit PCA on the embeddings HxW F-dimensional feature vectors and convert the resulting dimensions into RGB space. For K-means, we set $K = 3$, which usually divides the image into objects, people and background.

Our system is by no means perfect. We show failure modes in Figure 4. In addition to standard mispredictions by deep learning systems, some failures are intrinsic to the signal we use: if the object is unlikely to move, then our system is unlikely to see it as a held object.

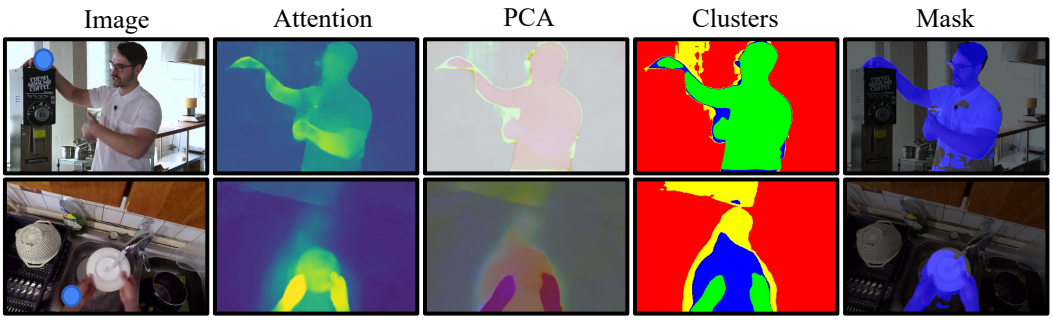

Figure 4: Failure cases of responsibility maps. Row 1 depicts a 100DOH example, where (arguably correctly), the responsibility prediction (nor the embedding) see the static coffee maker as a held object. Instead, the embedding is similar to that of the background. Row 2 depicts an EpicKitchens example, where despite a query image with a provided hand location, the responsibility prediction encompasses both hands, instead of just one hand and the object.

5 Discussion

In this work, we present a system for learning object embeddings in a relatively unsupervised manner. These embeddings allow us to distinguish an object from the background and people, but not from one another. A future work or next step of this research should be learning to separate these object instances from one another, such that embeddings are distinct based on the item present's class.

Our system works in scenes where there are 1 or more people present. It produces interesting split responsibilities when two hands are both touching an item, and it successfully identifies when people are touching people versus other objects. Occasionally, it erroneously extends responsibility predictions to adjacent items that are potentially just brushed by a moving hand, but overall it is quite capable of detecting what is actually held and directly contacted. We train and test on diverse datasets, filled with long tails of objects classes [38], yet the system generally performs well, leveraging the capability of the computed responsibilities pseudomasks to identify held object extents using only optical flow at training time.

Generally, if one visualizes segmentation results without image overlays, one is left with a very limited impression of the classes and extents of the world. People, cars, and other common objects are well segmented, yet the multifarious held objects that people interact with are often forgotten. We often neglect the necessity for understanding these uncommon object segments, yet a held object can have a very large impact on the reality of a scene. A person carrying a delicate glass vase needs a larger berth than someone holding a lunch box. We see our system as a first step for embodied agents to leverage hands a bridge to other physical objects, of an unknown class. This first step provides the community with a means to acquire weak-labels for segmentations of long-tailed object classes, supporting systems that may one day segment all objects.

Societal Impact: The world is filled with objects which do not fall into ImageNet's 1000 classes. Computer vision research generally does not deal with these objects, nor does it propose methods that reliably segment them. Systems that can learn about objects from video have the potential to enable many important positive downstream applications. This can range from helping understand interactions to better train systems to assist people.

On the other hand, our system can have negative downstream consequences. First, while the data in public videos is often closer to reality than Flickr images, deep learning systems can replicate and amplify the biases observed in these videos. While there has been a lot of effort twoards making datasets more reflective of everybody, there is still a lot of progress to be made. This requires both careful reflection during data curation and monitoring during deployment. Second, premature deployment of these systems without proper oversight can have substantial negative consequences, particularly for people not well-represented. Finally, while we are motivated by the high level goal of helping systems understand the physical world in order to assist humans (e.g., in robotics), our system for understanding what humans are doing with their hands can have of course be used for surveillance purposes.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

[3] Sven Bambach, Stefan Lee, David Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, 2015.

[4] 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.

[5] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *The European Conference on Computer Vision (ECCV)*, August 2020.

[6] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *European conference on computer vision*, pages 282–295. Springer, 2010.

[7] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. *arXiv preprint arXiv:2012.09856*, 2020.

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

[9] 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.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[11] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting anything that moves. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[12] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David Crandall. Hope-net: A graph-based model for hand-object pose estimation, 2020.

[13] 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.

[14] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[15] J. Gibson. *The ecological approach to visual perception*. Boston: Houghton Mifflin, 1979.

[16] Georgia Gkioxari, Ross Girshick, Piotr Dollar, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018.

[17] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 2020.

[18] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[21] Vladimir Iglovikov.

[22] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. *arXiv preprint arXiv:2006.14613*, 2020.

[23] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2020.

[24] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *proceedings of the European Conference on Computer Vision (ECCV)*, pages 207–223, 2018.

[25] Yin Li, Manohar Paluri, James M Rehg, and Piotr Dollár. Unsupervised learning of edges. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1619–1627, 2016.

[26] 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.

[27] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *arXiv preprint arXiv:1807.03247*, 2018.

[28] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *arXiv preprint arXiv:1606.00061*, 2016.

[29] A. Mittal, A. Zisserman, and P. H. S. Torr. Hand detection using multiple proposals. In *BMVC*, 2011.

[30] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai. Detecting hands and recognizing physical contact in the wild. 2020.

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

[32] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2701–2710, 2017.

[33] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[34] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017.

[35] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[38] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9869–9878, 2020.

[39] Jianbo Shi and Jitendra Malik. Motion segmentation and tracking using normalized cuts. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 1154–1160. IEEE, 1998.

[40] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artif. Life*, 11(1–2):13–30, Jan. 2005.

[41] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020.

[42] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. *arXiv preprint arXiv:2102.06191*, 2021.

[43] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019.

[44] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020.

[45] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019.

[46] Max Wertheimer. Laws of organization in perceptual forms. 1938.

[47] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. *arXiv preprint arXiv:2103.17263*, 2021.

[48] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[49] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. Lego: Learning edge with geometry all at once by watching videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 225–234, 2018.

[50] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] There are detailed experimental details in both the main paper and supplement.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Detailed training details are in both the main paper and supplement.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Our experiments are too costly to run to compute error bars without taking weeks of compute.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] + [No] EPIC-Kitchens obtained consent since they hired people to collect the data. 100DOH is a large Internet-scale dataset and so obtaining consent from each user is impractical. We are aware of the concerns with large-scale Internet video, and have therefore focused on also making sure our system is validated on a dataset like EPIC-Kitchens.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] These appear in the supplemental
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] Estimating hourly wage is difficult since we hired a crowdsourcing company, but more information appears in the supplement.