Self-Supervised Cross-View Correspondence with Predictive Cycle Consistency

Anonymous CVPR submission

Paper ID 13577

Abstract

Learning self-supervised visual correspondence is a long-studied task fundamental to visual understanding and human perception. However, existing correspondence methods largely focus on small image transformations, such as object tracking in high-framerate videos or learning pixel-to-pixel mappings between images with high view overlap. This severely limits their application in dynamic multi-view settings such as robot imitation learning. In this work, we introduce Predictive Cycle Consistency for learning object correspondence between extremely disjoint views of a scene without paired segmentation data. Our technique bootstraps object correspondence pseudolabels from raw image segmentations using conditional grayscale colorization and a cycle-consistency refinement prior. We then train deep ViTs on these pseudolabels, which we use to generate higher-quality pseudolabels and iteratively train better correspondence models. We demonstrate the performance of our method under both extreme in-the-wild camera view changes and across large temporal gaps in video. Our approach beats all prior supervised and prior SoTA self-supervised correspondence models on the EgoExo4D correspondence benchmark (+6.7 IoU Exo Query) and the prior SoTA self-supervised methods SiamMAE and DINO V1&V2 on the DAVIS-2017 and LVOS datasets across large frame gaps.

1. Introduction

At the core of learning is the discovery of recurring patterns. This is the purpose of visual correspondence: given multiple inputs to a scene—whether different camera angles, video frames, or other perspectives—how can we determine which objects are the same and which are different?

In recent years, the introduction of large-scale datasets and powerful model architectures has led to strong results in learning visual correspondence without the need for costly labeled object pairings. However, existing self-supervised object correspondence methods [18, 28, 52, 54] have overwhelmingly focused on domains with continuous or small transformations, such as Video Object Segmentation (VOS) in continuous videos, or dense pixel-to-pixel methods which necessarily assume that object surfaces overlap between views. This limits the application of self-supervised correspondence to *discontinuous* inputs, such as the first-and-third person EgoExo4D dataset [16], where objects are small, frequently interacted with, and camera pose or depth changes between views are unavailable. Learning correspondence in these challenging, real-world domains is crucial for tasks such as robot imitation learning [42, 44], object reidentification [13, 46], and scene understanding [16]. For example, in robot imitation learning, visual correspondence allows a robot to map observed actions from a third-person perspective to its own first-person actions.

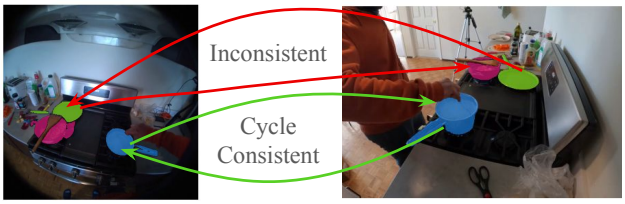

Figure 1. We extract cycle-consistent correspondence *at the object level* to guide self-supervision across *extreme view changes*

The fundamental challenge of learning visual correspondence across discontinuities is that a system must take into account an object and the surrounding environment holistically to answer the question of where *that* mug is instead of where *a* mug is. A particularly powerful approach for encouraging a holistic understanding in difficult data sources is predictive learning: training a model to de-corrupt a corrupted input. Predictive learning has shown strong results in a wide range of domains, such as natural language processing [9, 39], image classification [2, 21], and generative modeling [22]. Recently, predictive approaches have seen competitive results in object correspondence and representation learning through tasks such as image-conditioned grayscale coloration [52] and conditional masked autoencoders [18, 37].

However, existing predictive methods for correspon-

CVPR
#13577

Figure 2. Self-supervised object correspondence on EgoExo4D between Ego views (bottom row) and Exo views (top row) without any paired data. Objects with no match are not colored.

dence struggle to generalize effectively across challenging scenarios (Table 1). State-of-the-art approaches typically rely on K-Nearest-Neighbor matching of emergent learned representations, identifying corresponding regions by comparing the most similar model embeddings across views. Although the pretraining objectives of these methods promote spatial awareness within a scene, the resulting representations often entangle semantic object information with the spatial cues necessary for correspondence. This entanglement reduces robustness against semantically similar distractor objects, limiting generalization and performance in more demanding scenarios.

An effective approach for encouraging a spatially-consistent mapping between views is cycle consistency, which leverages the fact that correspondence is generally invertible across time and view changes [47, 54, 62]. Objects that can be tracked forward in a video can also be tracked backward; camera view changes can be undone. Existing applications of cycle consistency bootstrap long-term correspondence relations via local techniques like optical flow or video palindromes. However, these techniques rely on continuous input data such as high-framerate videos and require hand-crafted biases such as local neighborhoods to restrict a large search space. As a result, cycle consistency approaches have fallen out of favor compared to more modern approaches in challenging scenarios.

In this work, we propose Predictive Cycle Consistency (PCC) for the task of self-supervised object correspondence under extreme viewpoint changes in space (camera angle changes) and time (gaps between video frames). Our approach extends cycle-consistency to operate at the *object level* rather than focusing on features or image patches, allowing for robust correspondence in challenging scenarios. We do this by building on existing predictive approaches [18] and [52] to train a directional correspondence model on the simple, asymmetric task of grayscale colorization of a target image conditioned on a colorful source view. Using this, we extract object pairs that correspond to each other when running the correspondence model in *both* directions–source to target and target to source–to generate high-quality paired pseudolabels. We then train deep ViTs on these pseudolabels, which we use to generate higher-quality pseudolabels and iteratively train better correspondence models. All together, our approach combines the power of deep transformers trained with a predictive pretext task for scene understanding with a method to extract refined object boundaries.

Our approach learns object correspondences across a wide range of challenging scenarios, such as matching highly occluded objects (apron and hands, Fig. 2 left) and distinguishing between semantically similar objects (pieces of paper from a COVID test, Fig. 2 middle). We validate our method on several object correspondence benchmarks, achieving superior performance over all previous *labeled* approaches and state-of-the-art unlabeled correspondence methods on the EgoExo4D Object correspondence benchmark [16]. Additionally, our approach sets a new state of the art on widely used video tracking datasets DAVIS-2017 [38] and LVOS [23] under high viewpoint-change conditions. We commit to making our code and checkpoints open source upon acceptance. Our contributions are as follows:

1. We introduce Predictive Cycle Consistency, a technique that combines the powerful representation learning of predictive approaches with the refinement of cycle-consistency for self-supervised correspondence.
2. We propose a pseudo-labeling method that incorporates cycle-consistency on top of existing correspondence models to iteratively refine self-supervised object correspondence outputs.
3. We obtain state-of-the-art results on a suite of correspondence tasks from EgoExo4D, DAVIS, and LVOS.

2. Related Work

Dense Visual Correspondence. The visual correspondence task is fundamental to human visual perception and has a long established history in computer vision. Early techniques for correspondence focused on dense pixel-to-pixel correspondence using classical techniques such as optical flow [3, 24, 27, 30, 35, 48, 57]. In recent years, dense correspondence learning has advanced with deep learning methods that address challenging conditions, including significant viewpoint changes. However, existing dense methods for correspondence either require large amounts of costly labeled data to train, use camera pose [53], depth information [11, 41], or extract annotations through simulated data [50, 51]. Furthermore, implicit to dense correspondence is a restriction on allowed view changes: different views must contain overlapping object surfaces. As a result, our work focuses on the more semantically grounded task of

object level correspondence given segmentation masks.

Self-Supervised Object Correspondence Human vision excels at establishing visual correspondences across space and time without direct supervision, even in the face of occlusions, distractor objects, and object transformations. Inspired by this capability, much work has focused on self-supervised learning of object correspondence through video and scene data.

Contrastive approaches [20, 61] for object tracking learn correspondence by creating ground-truth annotations through applying strong image transformations. This powerful inductive bias has been extraordinarily effective in tracking, but these approaches are criticized for the large amount of engineering and hand-crafting necessary to create robust augmentations. *Cycle-consistency* [47, 54, 62] emerged as an early technique for learning deep self-supervised object correspondence by leveraging the invertibility of object correspondence through techniques such as training on video palindromes. However, cycle-consistency as the sole training objective has struggled on its own to maintain stability over long videos or through occlusions, causing it to be replaced by more powerful representation learning techniques [26, 61]. *Tracking-by-Matching* approaches are a recent introduction to learning zero-shot video instance segmentation. These approaches build on top of supervised image segmentation models [56] to create object-level representations. These representations either transfer from labeled data across domains [7] or are learned through maintaining invariance under hand-crafted augmentations [33, 34].

Visual Representation Learning for Correspondence Current state-of-the-art approaches for both supervised and unsupervised correspondence learning rely on strong base models to construct clean visual representations. In the past few years, these base models have moved away from ResNets [19] towards larger models such as Vision Transformers (ViTs) [10] trained using using self-supervised representation learning strategies on large datasets. For example, the DINO model family [4, 8, 36] emphasizes non-contrastive invariance under augmentation. Other methods use exponentially moving average models [1, 17] or siamese networks [5]. In cases involving multiple views, large-scale contrastive models effectively align embedding spaces across domains, achieving strong results in view-invariant learning and language-image matching [14, 25, 40, 58].

Masked Autoencoding [21] trains high-quality visual representations by randomly masking patches of an image at a high ratio and reconstructing the missing areas. A recent advancement in this approach for correspondence learning is SiamMAE [18], which adapts the infilling task by asymmetrically masking a future frame and predicting it based on an unmasked past frame. This approach, also used in cross-view masked completion models [59, 60], has achieved state-of-the-art performance in generating robust representations for object correspondence. In this work, we extend existing masked autoencoder architectures to enhance correspondence across extreme viewpoint changes.

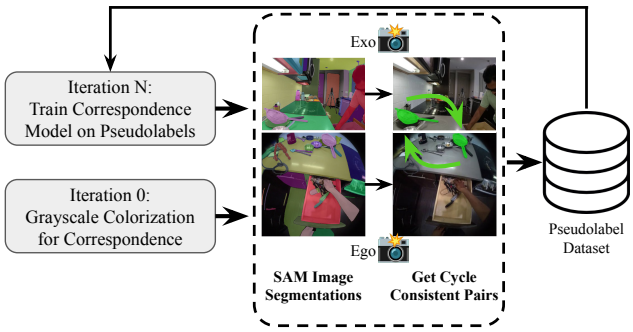

Figure 3. **High-Level Approach**. We 1) Enumerate all objects in each view, 2) Create pseudolabels that are cycle-consistent according to a previous correspondence model iteration, and 3) Use these pseudolabels to train a new correspondence model.

3. Approach

We tackle the task of discovering object correspondence between two images $\mathcal{I}_1 \in \mathbb{R}^{3\times H_1\times W_1}$ and $\mathcal{I}_2 \in \mathbb{R}^{3\times H_2\times W_2}$ of the same spatiotemporal scene. Specifically, we want to create a model takes in a binary segmentation mask $\mathcal{O}_1$ of an object in image $\mathcal{I}_1$ and outputs $\mathcal{O}_2$, the corresponding object segmentation mask in $\mathcal{I}_2$, or indicates that no correspondence exists.

At a high level, our correspondence pipeline (Fig. 3) operates by 1) Enumerating all objects in each image with existing segmentation models, 2) Using our proposed PCC technique to discover a subset of corresponding objects in each view, 3) Generating correspondence pseduolabels with these mined corresondences, and 4) Iteratively training new models and generating new pseudo labels to learn high-quality self-supervised visual correspondence.

3.1. Image Segmentation

We break down finding corresponding objects between two images into two subtasks: 1) Enumerate all objects in each image (Object detection and Image Segmentation) and 2) Discover which of these objects correspond to one another (Correspondence). This cascaded approach to correspondence has recently obtained state-of-the-art results in video object segmentation with low data or textual object descriptions [7, 34]. There is a great amount of prior work in object detection and image segmentation, with popular approaches being SAM [32] in the supervised setting and UnSAM [56] and CutLER [55] in the self-supervised setting. We follow prior object-centric correspondence work [7, 33, 34] and use SAM to segment an image into distinct object regions,

although we suspect self-supervised approaches could be applied for a completely unsupervised pipeline. The rest of this paper focuses on the task of (2), learning object correspondence without labels given object segmentations.

3.2. Grayscale Colorization

Grayscale colorization [52] was an early yet remarkably effective method for self-supervised object tracking in continuous videos. However, as computational resources grew and advanced architectures like Vision Transformers (ViTs) became prevalent, grayscale colorization fell out of favor, replaced by more complex and general pretext tasks such as masked autoencoding [12, 18, 21, 29, 37, 49]. To address the challenge of learning correspondence under extreme discontinuities, we reinvestigate the grayscale colorization approach and find that it reemerges as strikingly successful technique.

Our architecture for grayscale colorization is depicted in Figure 4 (a). Conditioned on our source image $\mathcal{I}_1$ we optimize to reconstruct $\mathcal{I}_2$ given its grayscale counterpart $\mathcal{I}_2^g$. We adapt our architecture from prior correspondence work [18, 60] and use a two-stage ViT [10] that 1) passes $\mathcal{I}_1$ and $\mathcal{I}_2$ separately through an encoder module $\mathbf{E}$ to obtain $\mathbf{E}(\mathcal{I}_1), \mathbf{E}(\mathcal{I}_2^g)$, and 2) passes the encoded grayscale image $\mathbf{E}(\mathcal{I}_2^g)$ into a decoder $\mathbf{D}$ that merges the colorful image conditioning using cross attention: $\mathbf{D}(\mathbf{E}(\mathcal{I}_1), \mathbf{E}(\mathcal{I}_2^g))$. For optimization, we use a simple mean-squared-error loss in the RGB color space.

3.3. Visual Correspondence from Colorization

We now propose an approach to extract object correspondence from the inputs and outputs of conditional colorization models based on a simple observation: if we augment an object's color in the source view, we would expect the output of the colorization model to change the object's color in the target view as well.

Our approach is demonstrated in Figure 5. Given a binary object segmentation map $\mathcal{O}_1$ in $\mathcal{I}_1$ we seek to obtain the corresponding segmentation mask $\mathcal{O}_2$ in view $\mathcal{I}_2$. We perform two forward passes of our grayscale colorization model, $\mathbf{F}$. First, we perform grayscale cross-view infilling of $\mathcal{I}_2^g$ given a colorized $\mathcal{I}_1$ from before: $y = \mathbf{F}(\mathcal{I}_2^g, \mathcal{I}_1)$. For our second model pass, we augment $\mathcal{I}_1$ at the location of the segmentation mask $\mathcal{O}_1$ by adding a constant color offset to each channel. We denote this augmented image as $\mathcal{I}_1'$, and the RGB augmentation as a vector $c \in \mathbb{R}^3$. We then colorize $\mathcal{I}_2^g$ conditioned on this augmented image $y' = \mathbf{F}(\mathcal{I}_2^g, \mathcal{I}_1')$. To identify where objects correspond, we then take the average absolute difference between y and y' over each color channel and normalize to sum to one, outputting an object segmentation heatmap $\mathcal{H} \in \mathbb{R}_+^{H_2 \times W_2}$:

$$\mathcal{H}_{ij} = \frac{\sum_{c=1}^{3} |\mathbf{F}(\mathcal{I}_2^g, \mathcal{I}_1)_{cij} - \mathbf{F}(\mathcal{I}_2^g, \mathcal{I}_1')_{cij}|}{\sum_{c,k,l} |\mathbf{F}(\mathcal{I}_2^g, \mathcal{I}_1)_{ckl} - \mathbf{F}(\mathcal{I}_2^g, \mathcal{I}_1')_{ckl}|} \quad (1)$$

Importantly, this approach is entirely blackbox with respect to the grayscale colorization model. As a result, the spatial awareness of objects captured during grayscale model pretraining naturally emerges during extraction. This is not necessarily true for masked autoencoder approaches like SiamMAE, where internal model representations are used to extract correspondences that may not sufficiently encode cross-view information to disentangle semantically similar distractors.

3.4. Generating Cycle-Consistent Correspondence

In this section, our goal is to create a set of pseudolabels containing corresponding objects between two images $\mathcal{I}_1$ and $\mathcal{I}_2$ given their respective object segmentations $\{\mathcal{O}_{1,i}\}_{i=1}^{N_1}$ and $\{\mathcal{O}_{2,j}\}_{j=1}^{N_2}$ output by SAM, where $\mathcal{I}_1, \mathcal{I}_2$ have N_1, N_2 segmented objects respectively. Following Sec. 3.3, we generate an augmented $\mathcal{I}_{1,i}'$ for each segmentation mask $\mathcal{O}_{1,i}$ and calculate the output correspondence heatmap in $\mathcal{I}_2$, which we write as $\mathcal{H}_i^{1\to2}$. Here, "$1 \to 2$" represents predicting $\mathcal{I}_2$ conditioned on $\mathcal{I}_1$. To reduce variance, we use the same augmentation vector c for each augmentation. By caching the unaugmented pass, calculating all $\mathcal{H}_i^{1\to2}$ takes $N_1 + 1$ forward passes.

We then define the similarity between $\mathcal{H}_i^{1\to2}$ and $\mathcal{O}_{2,j}$ to be the amount of weight $\mathcal{H}_i^{1\to2}$ places on the segmentation mask of $\mathcal{O}_{2,j}$, equal to the Frobenius inner product:

$$\text{Sim}(\mathcal{H}_i^{1\to2}, \mathcal{O}_{2,j}) = \langle \mathcal{H}_i^{1\to2}, \mathcal{O}_{2,j} \rangle_F \quad (2)$$

Then, we pair every object $\mathcal{O}_{2,j}$ with the most similar heatmap $\mathcal{H}_i^{1\to2}$, which we write as $\mathbf{P}_j^{1\to2}$:

$$\mathbf{P}_j^{1\to2} = \arg\max_i \ \text{Sim}(\mathcal{H}_i^{1\to2}, \mathcal{O}_{2,j}) \quad (3)$$

There is an important difference between taking the argmax over the source object augmentation heatmaps and taking the argmax over the objects in the target image. Outputs of grayscale colorization can be correlated with many factors such as object semantics (the sky is probably blue but a balloon can be any color) or lighting conditions, which may mean that some objects in the target view are more susceptible to having their predicted colors change with respect to augmented inputs than others. As a result, the object with the greatest color change in I_2 is often the same for many $H_i^{1\to2}$. By taking the most similar $H_i^{1\to2}$ for each $\mathcal{O}_{2,j}$, we normalize over the easiness of each object to be changed as the result of an augmentation in I_1 during grayscale colorization.

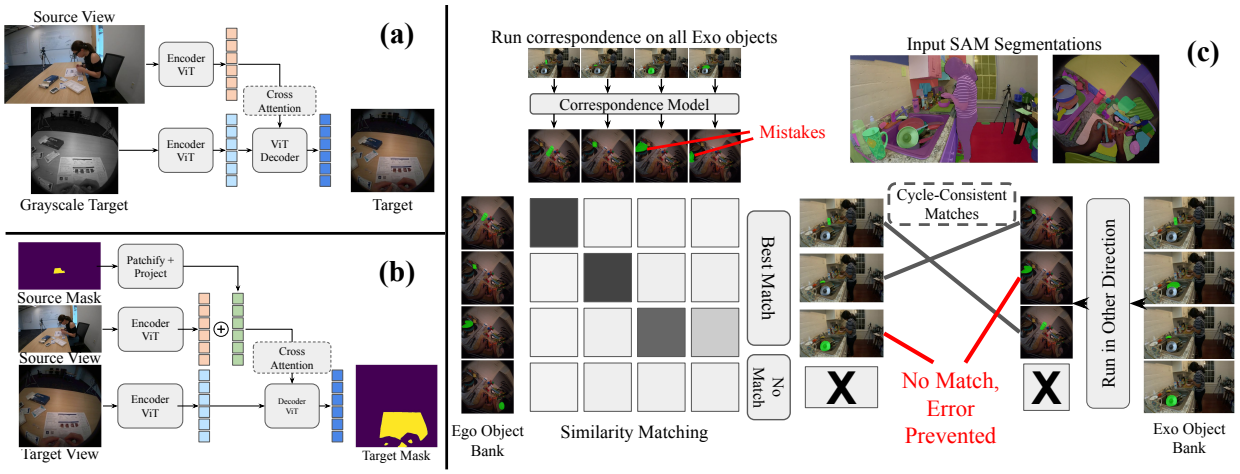

Figure 4. Model Architecture and Cycle Consistency. (a) Our architecture for grayscale color completion, using a shared Encoder for each view and a Decoder with Cross and Self Attention. (b) Our architecture for correspondence, which inserts a projection to represent the mask. (c) Our cycle-consistency pipeline. We run correspondence using the model from (a) or (b) on all objects in each view, find the best matching objects, and extract cycle-consistent matches.

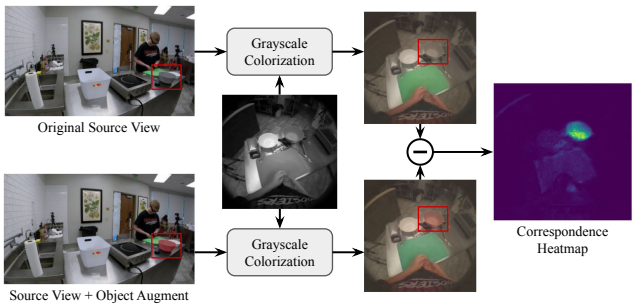

Figure 5. A simple method to extract object correspondence from a deep conditional grayscale colorization model.

We repeat the above process in the reverse direction to extract $\mathcal{H}_j^{2\to1}$ and $\mathbf{P}_i^{2\to1}$. Finally, we extract cycle-consistent pseudolabels at the object level by taking all pairs (i, j) where $P_j^{1\to2} = i$ and $P_i^{2\to1} = j$. This cycle consistency enforces that augmenting O_i in I_1 leads to changes at O_j in I_2 and augmenting O_j in I_2 leads to changes at O_i in I_1 during grayscale infilling, which strongly implies correspondence. We then use these extracted pseudolabels on each image to generate a dataset of labeled correspondence data with paired images and segmentation masks.

3.5. Correspondence Model

Our end goal is to have a correspondence model $\mathbf{G}$ than can take in a source image $\mathcal{I}_1$, source object segmentation mask $\mathcal{O}_1$, and target image $\mathcal{I}_2$ and output the corresponding object segmentation mask for the target image. To train such a model on our extracted pseudolabels or ground truth labeled data, we slightly modify our grayscale colorization architecture, shown in Fig 4 (b). We patchify the segmentation mask $\mathcal{O}_1$, apply a linear projection, and add the output to $\mathbf{E}(\mathcal{I}_1)$. The decoder then takes in the encoded target view $\mathbf{E}(\mathcal{I}_2)$ and uses cross attention to query the source view and source segmentation mask. As done in [16] we use a Dice Loss [45] and a BCE Loss to predict the target mask.

3.6. Iterative Pseudolabeling and Training

We repeat our pseudolabeling pipeline by substituting grayscale colorization with a trained correspondence model. Assuming that a correspondence model $\mathbf{G}$ outputs a binary mask, we replace the similarity score (Equation 2) with intersection over union (IoU) and make no other changes. We can then use these new pseudolabels to train new correspondence models, which in turn can generate new pseudolabels. In our experiments, we find that running three rounds of pseudolabeling (one grayscale, two iterative) saturates quality.

4. Experiments

For comprehensive evaluation, we train and evaluate our PCC pipeline from start to finish in two settings 1) high camera pose changes and 2) large temporal gaps in videos. We describe our training and evaluation approaches for each setting in Sec. 4.1 and present results in Sec. 4.2 and 4.3.

4.1. Experimental Setup

Datasets To train and evaluate the quality of PCC across extreme camera viewpoint changes, we use the EgoExo4D correspondence dataset [16]. EgoExo4D [16] builds on Ego4D [15] and captures over 1000 hours of complex human actions such as cooking and basketball from several

time-synchronized camera views. The camera views consist of one egocentric (ego) view captured by person performing the action and one or more exocentric (exo) camera views that capture the holistic action environment from the third person. We train and evaluate on the object correspondence benchmark in EgoExo4D, which is unique across correspondence datasets in that it focuses on extremely challenging settings with many small objects, high occlusion, and extreme camera angle differences in a dynamic environment. Correspondence is evaluated in two directions, either with taking a query mask in the ego view and segmenting the corresponding object in a paired exo view (ego query), or the inverse task from the exo to the ego view (exo query).

To run PCC on video object segmentation across large temporal gaps, we train on the commonly used Kinetics-400 dataset [31], consisting of large-scale action videos taken from the web. We evaluate temporal correspondence on DAVIS-2017 [38], which captures densely segmented short videos and LVOS [23], which measures video object segmentation across long dynamic videos with an average duration of over a minute.

Backbone We implement our correspondence models using a ViT-B with a patch size of 16 to match the compute and parameter count of current state-of-the-art self-supervised correspondence work [18, 37]. We initialize our model from CroCo v2 [59, 60], which pretrains with the task of image conditioned cross-view autoencoding.

Grayscale Model We train our grayscale colorization models separately for EgoExo4D and Kinetics-400. For EgoExo4D, for each ego view in the dataset, we 1) select a random corresponding exo camera, 2) choose a random synchronized frame from each view, and 3) choose with 50% probability whether to grayscale the ego or exo view. For Kinetics-400, we follow SiamMAE and select two random frames between 4 and 48 frames apart and colorize the grayscale future frame conditioned on the past frame. We train each grayscale colorization model for 60k updates with a batch size of 256, which corresponds to 200 epochs for EgoExo4D and 50 epochs for Kinetics-400. Additional hyperparameters are in Appendix 6.1.

Predictive Cycle Consistency We then run Predictive Cycle Consistency to extract paired pseudolabels. We use the ViT-H version of SAM to run image segmentation on video frames to extract $\{\mathcal{O}_{1,i}\}_{i=1}^{N_1}$ and $\{\mathcal{O}_{2,j}\}_{j=1}^{N_2}$. For EgoExo4D, we segment every labeled (ego,exo) frame pair in the correspondence benchmark, which generally amounts to one frame per second. For Kinetics-400, we segment two random frames per video selected 2 seconds apart, and we ablate the length between frames in our experiments.

Correspondence Model For a fair comparison, we match the compute of the EgoExo4D supervised baselines and train our correspondence models on 5M supervised examples, divided into 10k updates with a batch size of 512. We train for the same duration on Kinetics-400, and additional hyperparameters are provided in Appendix 6.1. For a given input example, we select select a random cycle-consistent pseudolabel as a training target, and with a 50% chance select whether to train with an ego or exo query. On 25% of inputs, we replace the correspondence objective with learning correspondence existence accuracy, where we select a uniformly random SAM image segment, label it positive if it is also a cycle-consistent pseudolabel and negative otherwise.

Baselines To the best of our knowledge, this is the first work to tackle self-supervised correspondence across the uniquely extreme viewpoint changes on EgoExo4D. As a result, we are rigorous about reimplementing the prior SoTA open source correspondence models. As demonstrated in the SiamMAE paper [18], the current best approaches for correspondence without labeled data are deep ViT models. In particular, SiamMAE makes two key observations 1) a small patch size (ViT-X/8) results in significantly stronger correspondence results (+9.5 on DAVIS-17 for ViT-S/8 vs SiamMAE ViT-S/16) and 2) despite deep work in prior correspondence approaches, DINO v1 [4], which focuses on learning robust and general visual representations at scale, outperforms all approaches except SiamMAE. As a result, for prior work we compare against 1) SiamMAE ViT-S/8, 2) SoTA DINO v1 and v2 models, and 3) CroCo v2 (to compare versus our baseline model architecture). For fairness, we continually pretrain SiamMAE and CroCO v2 on EgoExo4D. Additional implementation details are in Appendix 6. Furthermore, although our model does not use SAM at inference, we do use image segmentation for pseudolabel generation. To account for this, we implement a setting where we use SAM ViT-H to select the object with the highest IoU. For the DAVIS-17 and LVOS baselines, we similarly compare against the aforementioned models, however find that SAM did not help because it cut off objects (e.g. selecting a tire instead of a bike), which we discuss more in Appendix 6.

4.2. Results Across Space

We report quantitative results for PCC across extreme viewpoint changes on the EgoExo4D correspondence benchmark [16] in Table 1. Following the EgoExo4D baseline, we measure 1) IoU (denoted as $\mathcal{J}_m$ in VOS datasets), 2) Location Score, representing the mean-squared-error distance between the predicted and ground truth centroid, 3) Contour Accuracy (CA, denoted as $\mathcal{F}_m$ in VOS datasets), and 4) object presence balanced accuracy (Bal. Acc.). We use the official code for EgoExo4D for evaluation, and run all baselines at 480p, as done in both [16] and [18].

Our results demonstrate that PCC outperforms all prior labeled and all prior SoTA self-supervised approaches across all metrics. We observe that the DINO family

Table 1. Results on of the EgoExo4D [16] correspondence benchmark (v1, test set). ♢: Continual Pretraining on EgoExo4D ♣: Models with access to multiple frames per view. Bold is best, underlined is second best. Top labeled, middle and bottom self-supervised.

		Ego Query				Exo Query			
Method	**Backbone**	**Bal. Acc. ↑**	**IoU ↑**	**Loc. Score ↓**	**CA.↑**	**Bal. Acc.↑**	**IoU↑**	**Loc. Score↓**	**CA.↑**
XSegTx [16]	SegSwap [43] + ViT-B	62.63	13.88	0.154	0.239	74.6	21.8	0.133	0.265
XMem♣ [6]	ResNet-50 [19] + Memory	42.33	13.07	0.312	0.182	56.96	10.2	0.249	0.125
XView-XMem♣ [16]	XMem + ViT-B	53.28	22.14	0.176	0.325	59.36	23.56	0.186	0.308
XView-XMem (+ XSegTx)♣ [16]	XMem + SegSwa + ViT-B	54.61	22.5	0.139	0.347	52.28	19.39	0.208	0.255
Ours Supervised	ViT-B/16	74.7	38.41	0.037	**0.603**	**88.45**	43.70	**0.049**	0.555
Ours Supervised + PCC	ViT-B/16	**76.9**	**39.01**	**0.033**	0.600	87.23	**47.06**	0.054	**0.590**
SiamMAE♢ [18]	ViT-S/8	∅	12.24	0.170	0.185	∅	14.18	0.159	0.198
CrocoV2♢ [59]	ViT-B/16	∅	7.14	0.200	0.138	∅	9.56	0.164	0.136
DINO [4]	ViT-B/8	∅	12.55	0.153	0.178	∅	15.94	0.137	0.246
DINOv2+Registers [36] [8]	ViT-B/14	∅	20.26	0.125	0.299	∅	24.6	0.169	0.307
SiamMAE♢+SAM	ViT-S/8+ViT-H	∅	17.97	0.180	0.254	∅	24.05	0.143	0.315
DINOv2+Registers+SAM	ViT-B/14+ViT-H	∅	28.92	0.153	0.365	∅	34.78	0.123	0.433
Grayscale Coloration + SAM♢	ViT-B/16+ViT-H	∅	20.82	0.110	0.311	∅	19.50	0.109	0.276
PCC Iter 1	ViT-B/16	∅	26.41	0.085	0.396	∅	34.35	0.090	0.436
PCC Iter 2	ViT-B/16	65.22	**29.98**	**0.083**	**0.446**	66.40	40.41	0.079	0.502
PCC Iter 3	ViT-B/16	60.66	29.89	0.094	0.432	67.90	**41.45**	**0.071**	**0.508**

of models, which learns robustness against strong image augmentations, outperforms the predictive SiamMAE and CroCo V2 approaches. Although applying SAM at inference time strongly improves self-supervised approaches in IoU and Contour Accuracy, it harms spatial accuracy as measured by Location Score, where PCC dramatically outperforms all prior work. We find that our self-supervised method performs relatively better in the Exo Query setting, where input objects are small, but output objects are generally larger. We find that iteratively training PCC strongly improves results, although performance largely saturates after the second iteration.

We further apply our correspondence model architecture (Sec. 3.5) on the labeled training data split of EgoExo4D, achieving state-of-the-art results compared to prior supervised approaches. When combining the supervised data from EgoExo4D with PCC Iteration 3 pseudolabels, we observe improvements over the supervised baseline, particularly in the Exo Query setting, suggesting that PCC captures complementary information to the EgoExo4D labels.

We show qualitative results of PCC on EgoExo4D in 6. As shown in rows 3 and 4, our model can locate corresponding objects even under significant occlusions. However, it sometimes struggles to produce fine-grained masks. Examples of pseudolabels on EgoExo4D are visible in 2.

Figure 6. Qualitative results of PCC on the EgoExo4D correspondence benchmark.

4.3. Results Across Time

We present quantitative results for PCC across large temporal gaps between video frames on the DAVIS-17 dataset [38] in Table 2. We follow the evaluation methodology in [11] and evaluate correspondences on all video frame pairs in DAVIS with a temporal gap of 20 frames. Evaluation is restricted to only cover objects present in both views, and report $\mathcal{J}\&\mathcal{F}_m$, $\mathcal{J}_m$, and $\mathcal{F}_m$. We report quantitative results on LVOS in Figure 7. Enabled by the large lengths of LVOS videos, we report results across a range of temporal gaps between 5 frames (the LVOS annotation rate) and 400 frames. We again restrict evaluation to objects present in both views and report $\mathcal{J}\&\mathcal{F}_m$.

On DAVIS, we find that PCC significantly outperforms the prior SOTA approaches for VOS. Similar to the EgoExo setting, PCC shows strong improvements after the first iteration, but performance saturates after the second. On LVOS, although SiamMAE achieves SoTA for temporal gaps of 5 and 10 frames, PCC shows comparatively better

Table 2. SoTA Comparison of Video Object Correspondence Methods on DAVIS-17 [38] Val with a temporal gap of 20 frames

Method	$\mathcal{J}\&\mathcal{F}_m\uparrow$	$\mathcal{J}_m\uparrow$	$\mathcal{F}_m$
SiamMAE [18]	60.7	58.4	62.9
CrocoV2 + Cont. Pretrain [59]	40.0	37.4	42.5
DINO ViTs/8 [4]	64.5	61.6	67.5
DINO ViTb/8 [4]	66.4	63.7	69.2
DINOv2 + Reg ViTb/14 [8]	62.1	59.6	64.8
PCC Iter 1	64.4	61.3	67.5
PCC Iter 2	69.7	67.0	72.4
PCC Iter 3	**70.2**	**67.8**	**72.7**

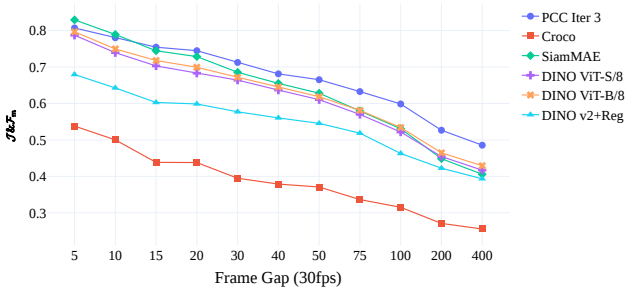

Figure 7. Results on LVOS-V1 Val at different temporal gaps.

performance as the frame gap increases, achieving SoTA results for all other temporal gaps. Interestingly, SiamMAE sees a sharper decline as the frame gap widens compared to DINO. We suspect this suggests that at shorter frame gaps, precise object segmentation is most critical, whereas at larger frame gaps, robustness to object semantics becomes more important.

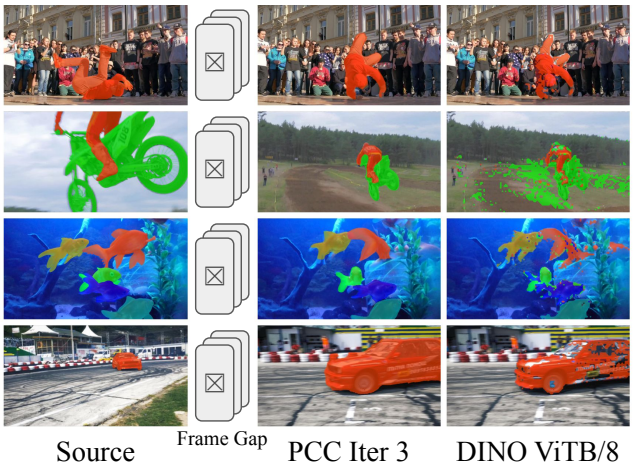

Figure 8. Qualitative results on correspondence across time. Representational approaches degrade quickly with larger frame gaps.

We demonstrate qualitative results for object recognition across temporal gaps in 8. We observe that representational approaches such as DINO struggle to maintain object consistency (bottom row), handle semantically similar distractors with moderate movement (third row), and manage heavy object occlusions (second row). In contrast, PCC maintains high-quality correspondence throughout.

4.4. Ablations

Table 3. **Ablation: Model configuration** Supervised Training, EgoExo4D Validation Set. Grayscale continued training harms supervised training. A deeper decoder improves performance.

Pretraining	Dec. Params.	Ego/Exo Query IoU ↑	Ego/Exo Query Loc. Score ↓	Ego/Exo Query CA. ↑
MAE	50M	36.9/40.2	0.043/0.078	0.57/0.51
MAE	100M	**37.7/43.7**	**0.042/0.060**	**0.60/0.56**
Grayscale	100M	31.0/34.3	0.048/0.089	0.49/0.43

Table 4. **Ablation: Frame Gap** Davis 2017 Validation Set. We ablate the framerate gap on Kinetics-400 during pseudolabel extraction on Iteration 3.

Pseudolabel Temporal Gap	$\mathcal{J}\&\mathcal{F}_m$ 2 Frames	10 Frames	30 Frames
2 seconds	**83.9**	75.8	65.4
4 seconds	83.5	**76.2**	66.3
6 seconds	83.7	**76.2**	**67.7**

We ablate our model configuration on supervised EgoExo4D correspondence in Table 3. Our results show that pretraining with a grayscale objective on EgoExo4D, rather than a MAE objective, significantly degrades performance. This suggests that the gains from PCC are not due to better representation learning, but rather arise from the full cycle-consistent pipeline.

We investigate the impact of the temporal gap between images for pseudolabeling PCC on Kinetics-400 in Table 4. We find that training with a longer gap improves longer-term correspondence.

5. Conclusion

In this work, we address the task of learning self-supervised visual correspondence across significant camera view shifts and large temporal gaps in video. Our approach, *Predictive Cycle Consistency*, achieves state-of-the-art performance on the EgoExo4D correspondence benchmark and in low-frame-rate scenarios on the DAVIS-17 and LVOS datasets. Furthermore, we show that applying cycle consistency to correspondence predictions *at the object level* is an strikingly effective strategy for generating clean and consistent correspondence pseudolabels. Overall, our work extends the capabilities of self-supervised learning to establish correspondence in highly disjoint and challenging domains, paving the way for improved understanding of dynamic scenes in the natural world.

References

[1] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 3

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1

[3] Steven S. Beauchemin and John L. Barron. The computation of optical flow. *ACM computing surveys (CSUR)*, 27(3):433–466, 1995. 2

[4] 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 International Conference on Computer Vision (ICCV)*, 2021. 3, 6, 7, 8

[5] 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. 3

[6] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 7

[7] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023. 3, 2

[8] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023. 3, 7, 8

[9] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4

[11] Mohamed El Banani, Ignacio Rocco, David Novotny, Andrea Vedaldi, Natalia Neverova, Justin Johnson, and Ben Graham. Self-supervised correspondence estimation via multiview registration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1216–1225, 2023. 2, 7

[12] Alexandre Eymaël, Renaud Vandeghen, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Efficient image pre-training with siamese cropped masked autoencoders. *arXiv:2403.17823*, 2024. 4

[13] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 1

[14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 3

[15] 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. 5

[16] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 1, 2, 5, 6, 7

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[18] Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. Siamese masked autoencoders. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 1, 2, 3, 4, 6, 7, 8

[19] 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. 3, 7

[20] 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. 3

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 3, 4

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[23] Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13480–13492, 2023. 2, 6

[24] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2

[25] Alexandros Iosifidis, Anastasios Tefas, and Ioannis Pitas. View-invariant action recognition based on artificial neural networks. *IEEE transactions on neural networks and learning systems*, 23(3):412–424, 2012. 3

[26] 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. 3

[27] Joel Janai, Fatma Guney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 690–706, 2018. 2

[28] Zhenyu Jiang, Hanwen Jiang, and Yuke Zhu. Doduo: Dense visual correspondence from unsupervised semantic-aware flow. In *arXiv preprint arXiv:2309.15110*, 2023. 1

[29] Zhouqiang Jiang, Bowen Wang, Tong Xiang, Zhaofeng Niu, Hong Tang, Guangshun Li, and Liangzhi Li. Concatenated masked autoencoders as spatial-temporal learner. *arXiv preprint arXiv:2311.00961*, 2023. 4, 1

[30] Rico Jonschkowski, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 557–572. Springer, 2020. 2

[31] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6, 1

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3

[33] Liulei Li, Wenguan Wang, Tianfei Zhou, Jianwu Li, and Yi Yang. Unified mask embedding and correspondence learning for self-supervised video segmentation. In *CVPR*, 2023. 3, 2

[34] Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. *CVPR*, 2024. 3, 2

[35] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4571–4580, 2019. 2

[36] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 3, 7, 1

[37] Gensheng Pei, Tao Chen, Xiruo Jiang, Huafeng Liu, Zeren Sun, and Yazhou Yao. Videomac: Video masked autoencoders meet convnets. In *CVPR*, 2024. 1, 4, 6

[38] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 2, 6, 7, 8, 1

[39] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 1

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[41] Tanner Schmidt, Richard Newcombe, and Dieter Fox. Self-supervised visual descriptor learning for dense correspondence. *IEEE Robotics and Automation Letters*, 2(2):420–427, 2017. 2

[42] Jinghuan Shang and Michael S Ryoo. Self-supervised disentangled representation learning for third-person imitation learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 214–221. IEEE, 2021. 1

[43] Xi Shen, Alexei A Efros, Armand Joulin, and Mathieu Aubry. Learning co-segmentation by segment swapping for retrieval and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5082–5092, 2022. 7

[44] Bradly C Stadie, Pieter Abbeel, and Ilya Sutskever. Third-person imitation learning. *arXiv preprint arXiv:1703.01703*, 2017. 1

[45] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 5

[46] Hao Tang, Kevin J Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. Egotracks: A long-term egocentric visual object tracking dataset. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[47] Yansong Tang, Zhenyu Jiang, Zhenda Xie, Yue Cao, Zheng Zhang, Philip HS Torr, and Han Hu. Breaking shortcut: Exploring fully convolutional cycle-consistency for video correspondence learning. *arXiv preprint arXiv:2105.05838*, 2021. 2, 3

[48] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[49] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 4

[50] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10346–10356, 2021. 2

[51] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Probabilistic warp consistency for weakly-supervised semantic correspondences. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8708–8718, 2022. 2

[52] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018. 1, 2, 4

[53] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 757–774. Springer, 2020. 2

[54] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019. 1, 2, 3

[55] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3124–3134, 2023. 3

[56] XuDong Wang, Jingfeng Yang, and Trevor Darrell. Segment anything without supervision. *NeurIPS 2024*, 2024. 3

[57] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4884–4893, 2018. 2

[58] Daniel Weinland, Mustafa Özuysal, and Pascal Fua. Making action recognition robust to occlusions and viewpoint changes. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III 11*, pages 635–648. Springer, 2010. 3

[59] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. In *ICCV*, 2023. 3, 6, 7, 8, 1

[60] Weinzaepfel, Philippe and Leroy, Vincent and Lucas, Thomas and Brégier, Romain and Cabon, Yohann and Arora, Vaibhav and Antsfeld, Leonid and Chidlovskii, Boris and Csurka, Gabriela and Revaud Jérôme. CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion. In *NeurIPS*, 2022. 3, 4, 6

[61] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10075–10085, 2021. 3

[62] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 117–126, 2016. 2, 3

Self-Supervised Cross-View Correspondence with Predictive Cycle Consistency

Supplementary Material

6. Implementation Details

In this section, we detail the specific setups for the following components:

1. Our Grayscale Colorization model (Section 3.2),
2. Our ViT Correspondence models (Section 3.5),
3. Our implementation of predictive approaches SiamMAE [18] and CroCoV2 [59], and
4. Our parameterization of SAM (Segment Anything Model) for tracking.

6.1. Hyperparameters

6.1.1 Grayscale and Correspondence Model

We implement our Grayscale Colorization model and Correspondence Model using the CroCoV2 [59] base architecture. Starting from the CroCoV2 Base-Decoder checkpoint, we continue pretraining with either the Grayscale Colorization objective or the original Cross-View MAE objective from CroCoV2 on datasets such as EgoExo4D [16] or Kinetics-400 [31]. The hyperparameters for continued pretraining are listed in Table 5.

Table 5. Hyperparameters for Grayscale Colorization and Correspondence Model Continual Training.

	Grayscale Colorization (Sec. 3.2)	Correspondence Model (Sec. 3.5)
Encoder Layers	12	12
Encoder Embed Dim	768	768
Decoder Layers	12	12
Decoder Embed Dim	768	768
MLP Dim	3072	3072
Learning rate	1.5×10^{-4}	1.5×10^{-4}
Adam β_1 / β_2	0.9 / 0.98	0.9 / 0.98
Weight decay	0.01	0.01
Learning rate schedule	Linear Decay	Linear Decay
Dropout	0.1	0.1
Warmup updates	8,000	8,000
Batch size	256	256
Updates	60,000	60,000
Training Objective	Colorization (RBG MSE Loss)	MAE
Kinetics-400 Time Gap	4-48 Frames	4-48 Frames

This continued pretraining results in the Grayscale Colorization model that we use to initialize PCC, extracting correspondence with the technique in Sec. 3.3. For our final PCC Correspondence Model, we further train using PCC pseudolabels as described in Section 3.5. Table 6 outlines the hyperparameters used for this additional training.

6.1.2 Baseline Implementation

To ensure fairness during evaluation, we continually pretrain CroCoV2 [59] and SiamMAE [18] on EgoExo4D [16] before measuring correspondence. The settings for continually pretraining CroCoV2 are outlined in Section 6.1.

Table 6. PCC Correspondence Model Hyperparameters. For each domain (EgoExo4D or Kinetics-400) we initialize our PCC Correspondence Model parameters with a continually pretrained MAE (Table 5)

	EgoExo4D [16] Correspondence	Kinetics-400 [16] Correspondence
Encoder Layers	12	12
Encoder Embed Dim	768	768
Decoder Layers	12	12
Decoder Embed Dim	768	768
MLP Dim	3072	3072
Learning rate	1.5×10^{-4}	1.5×10^{-4}
Adam β_1 / β_2	0.9 / 0.98	0.9 / 0.98
Weight decay	0.01	0.01
Learning rate schedule	Linear Decay	Linear Decay
Dropout	0.1	0.1
Warmup updates	2,000	2,000
Batch size	256	256
Updates	10,000	10,000
Training Objective	DICE + BCE	DICE + BCE
Kinetics-400 Time Gap	-	60 Frame Gap (2 sec)
EgoExo Parameters	50/50 Ego→Exo/Exo→Ego	-
Image size	240x240 (Ego) 240x416 (Exo)	224x224

Since the SiamMAE [18] code and checkpoints are not publicly available, we reimplement their approach by adapting the published CAT-MAE [29] codebase and checkpoints. We continually train SiamMAE using the CAT-MAE hyperparameters on Kinetics-400 for 60,000 steps with a batch size of 256. To validate our reimplementation, we evaluate our model on the DAVIS-2017 validation set [38], achieving a $\mathcal{J}\&\mathcal{F}_m$ score of 70.6, closely matching the original SiamMAE score of 71.4. For EgoExo4D, we continually pretrain this checkpoint for an additional 60000 steps at a batch size of 256, otherwise using the same settings.

We exclude DINO [36] models from continual pretraining on EgoExo4D due to a lack of diversity of data for image augmentation (EgoExo4D only has 123 unique sites used for data collection). Additionally, models employing exponentially moving average teachers require extensive tuning of the moving average temperature, making continual pretraining more challenging.

For our baselines, we adapt the K-Nearest-Neighbor implementation from [54]. While originally designed for multiple video frames, we modify it to treat all evaluation scenarios as two-frame videos. The algorithm inherently supports different resolutions for the first frame query and subsequent frames, accommodating the differing aspect ratios of Ego-view and Exo-view images. As detailed in Section 4.2, we resample all Ego and Exo videos to have a minimum resolution of 480p and perform a grid search to optimize the parameter k and the temperature. For EgoExo4D evaluation, we omit the neighborhood size parameter, as there is no spatial continuity between Ego and Exo views.

CVPR
#13577

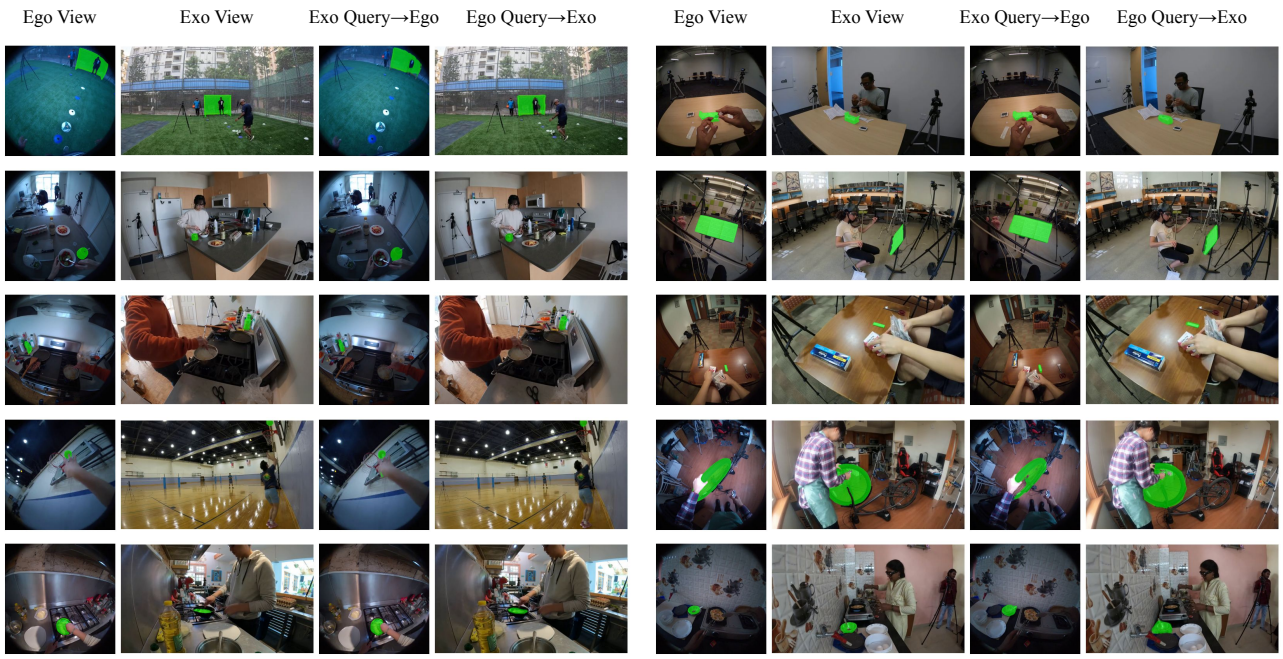

Figure 9. Additional Qualitative Results on the EgoExo4D Correspondence Benchmark.

However, for DAVIS-17 and LVOS evaluations, we independently grid search the neighborhood size parameter for each temporal distance.

6.1.3 SAM Configuration

To extract image segmentations from raw images in EgoExo4D, we use SAM with the standard point-grid prompting configuration, as demonstrated in [7, 33]. We note that this is different from the configuration of MASA [34], which uses bounding boxes extracted from an off-the-shelf object detection model using textual object descriptions. Because SAM is traditionally run on third-person videos, we gridsearch the Predicted IoU Threshold (0.88) and the Stability Score Threshold to (0.94) to have the highest IoU with ground truth object segmentation masks from the EgoExo4D validation set.

CVPR
#13577

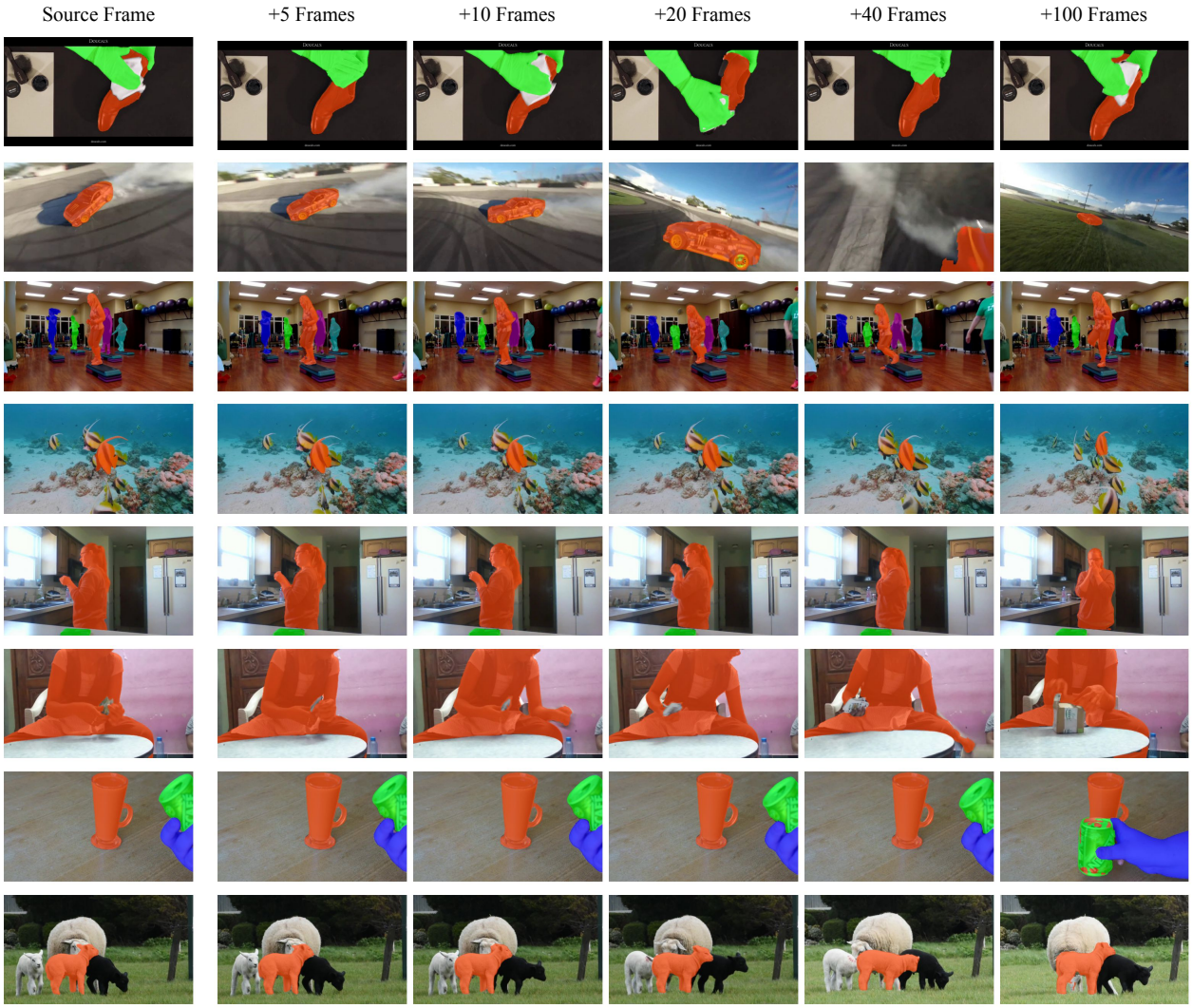

Figure 10. Additional Qualitative Results on LVOS with various frame gaps.