The Emergence of Objectness: Learning Zero-Shot Segmentation from Videos

Runtao Liu[1,2*] **Zhirong Wu**[1*] **Stella X. Yu**[3] **Stephen Lin**[1]

Microsoft Research Asia[1] John Hopkins University[2] UC Berkeley / ICSI[3]
runtao219@gmail.com stellayu@berkeley.edu {wuzhiron,stevelin}@microsoft.com

Abstract

Humans can easily segment moving objects without knowing what they are. That objectness could emerge from continuous visual observations motivates us to model grouping and movement concurrently from *unlabeled* videos. Our premise is that a video has different views of the same scene related by moving components, and the right region segmentation and region flow would allow mutual view synthesis which can be checked from the data itself without any external supervision.

Our model starts with two separate pathways: an appearance pathway that outputs feature-based region segmentation for a single image, and a motion pathway that outputs motion features for a pair of images. It then binds them in a conjoint representation called *segment flow* that pools flow offsets over each region and provides a gross characterization of moving regions for the entire scene. By training the model to minimize view synthesis errors based on segment flow, our appearance and motion pathways learn region segmentation and flow estimation automatically without building them up from low-level edges or optical flows respectively.

Our model demonstrates the surprising emergence of objectness in the appearance pathway, surpassing prior works on zero-shot object segmentation from an image, moving object segmentation from a video with unsupervised test-time adaptation, and semantic image segmentation by supervised fine-tuning. Our work is the first truly end-to-end zero-shot object segmentation from videos. It not only develops generic objectness for segmentation and tracking, but also outperforms prevalent image-based contrastive learning methods without augmentation engineering.

1 Introduction

In recent years, contrastive learning [1–3] has become a powerful model for obtaining high-level representations on generic images [4]. Despite their encouraging performance, contrastive models have two critical limitations. First, they heavily rely on hand-crafted image augmentations [5] to induce invariance, yet fail to account for complex variations such as object deformations and 3D viewpoints. Second, they still require additional labeled data and a fine-tuning stage for downstream applications, preventing use in a standalone fashion.

In this paper, we seek a zero-shot image model for detecting and segmenting objects by learning from unlabeled videos with minimal augmentation. As opposed to static images, a dynamic sequence of observations provides information about what is moving in the scene and how it moves. Such patterns not only reveal the boundary segment of an object, but also indicate hierarchical part organizations and even object semantics. Thus, as an easily accessible source of unlabeled data, videos provide rich natural supervision for learning image representations.

[*]Equal contribution. Work done when Runtao was a StarBridge intern at MSRA.

Preprint. Under review.

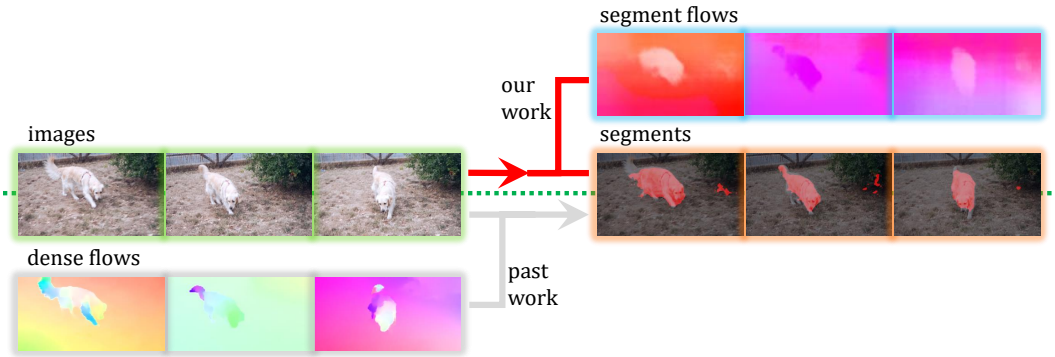

Figure 1: Our zero-shot object segmentation is learned from an unsupervised factorization of images into segments and their motions, whereas past work segments objects based on dense pixel-wise optical flows, which are brittle in the presence of noise, articulated movement, and abrupt motion.

A popular objective for self-supervised learning from videos is view synthesis. Concretely, given a source frame, a function is learned to warp the source frame to the target frame using photometric consistency as supervision. Dense optical flow [6] can be self-supervised in this manner. One could also learn a monocular depth network from videos with additional camera parameters [7]. The key idea is to find an appropriate representation which not only parameterizes the warping function, but also transfers to the target task. For example, the multi-plane image representation [8] is proposed to extrapolate between stereo pairs using view synthesis.

Unlike prior works that use view synthesis for low-level vision tasks, our goal is to tackle object segmentation which involves mid-level and high-level visual recognition. To this end, dense optical flow field which represents the low-level correspondence in a local manner would not suffice (see Figure 1). We therefore seek a new representation which could capture a gross characterization of moving regions for the entire scene. Deriving a representation explicitly for moving regions would allow the model to localize and segment objects.

Our approach decomposes view synthesis into two visual pathways: an appearance pathway to model "what is moving" by segmenting a static RGB image into separate regions, and a motion pathway to model "how it moves" by extracting motion features on a pair of images. The motion features are then used to predict flow offsets for individual regions assuming common fate [9] for all pixels within a region. The segment masks as well as their flow vectors jointly reconstruct a new representation called *segment flow*, which is used for view synthesis. In this way, object appearance and motion are decoupled, such that the appearance model for predicting segmentation would benefit from rich RGB signals. By conditioning on a region, the motion pathway is also tasked to solve a much simpler problem than dense flow. The two pathways are jointly trained with a reconstruction loss.

After self-supervised pretraining, we find that generic objectness detection and object segmentation automatically emerge from the model on one particular segment channel. Our model has the versatility for a variety of applications. First, the appearance pathway can be directly applied to novel images for dominant object segmentation in a zero-shot fashion. Second, it can also be fine-tuned for semantic segmentation on a small labeled dataset. Finally, with unsupervised test-time adaptation, the overall model can be transferred to novel videos for moving object segmentation without labels. Experimentally, we demonstrate strong performance on all of these applications, showing considerable improvements against the baselines.

The contributions of this work can be summarized as follows: 1) the first truly end-to-end zero-shot object segmentation from unlabeled videos; 2) a conceptually novel segment flow representation which goes beyond traditional dense optical flow; 3) a versatile model that can be applied to various image and video segmentation tasks. Our code is available at https://github.com/rt219/The-Emergence-of-Objectness.

2 Related Works

Video object segmentation. Segmentation of moving objects requires finding correspondences along the time dimension. A dominant line of work focuses on learning a representation for temporally

propagating segmentation masks. Such a representation may be learned with pixel-level object masks in videos with long-term relations [10, 11], or learned through self-supervision such as colorization [12] and cycle-consistency [13]. Given the annotation of object masks in the initial frame, the model tracks the object and propagates the segmentation through the remaining frames.

Fully unsupervised video object segmentation, without initial frame annotations, has received relatively little attention. NLC [14] and ARP [15] take a temporal clustering approach to this problem. Though they do not require segmentation annotations, elements of these algorithms depend on edge and saliency labels, and thus are not completely unsupervised. FTS [16] calculates the segmentation by obtaining a motion boundary from the optical flow map between frames. SAGE [17] takes into account multiple cues of edges, motion segmentation, and image saliency for video object segmentation. Contextual information separation [18] segments moving objects by exploiting the motion independence between the foreground and the background. A concurrent work based on motion grouping [19] clusters pixels with similar motion vectors. Both of these works rely on an off-the-shelf optical flow representation, which may be trained with [20, 21] or without [6] supervision.

Motion Segmentation. Classical methods for motion segmentation [22–24] cluster distinctive motion regions from the background based on two-frame optical flow. Supervised learning approaches [25, 26] map the optical flow field to segmentation masks. The requirement of dense and accurate optical flow may be problematic when the flow vectors are not smooth over time and vulnerable to articulated objects with inhomogeneous motion [27]. Motion has been shown to be effective when motion is considered over a large time interval [27]. An approach based on trajectory clustering [28] builds point trajectories over hundreds of frames, extracts descriptors for the point trajectories, and clusters them to obtain segmentation results. Though promising, such a global approach is computationally demanding. We turn our attention to modeling appearance on feature-based RGB representations, which provide rich cues (e.g. texture, color and edges) for perceptual organization, alleviating the need for dense pixel correspondence.

Unsupervised learning for segmentation. Human annotation of pixel-level segmentation is not only time-consuming, but also often inaccurate along object boundaries. Learning segmentation without labels is thus of great interest in practice. Segsort [29] predicts segmentation by learning to group super-pixels of similar appearance and context from static images. Later work [30] contrasts holistic mask proposals obtained from traditional bottom-up grouping.

A related line of work focuses on learning part segmentation from images and videos of the same object category, such as humans and faces. SCOPS [31] is a representative method learned in a self-supervised fashion. The general idea follows unsupervised landmark detection [32], where geometric invariance, representation equivariance and perceptual reconstructions are considered. Co-part segmentation [33] is also explored in videos, where motion provides a strong cue for part organization. A motion-supervised approach [34] models part motion between adjacent frames via affine parameters. Another work [35] implements a similar idea in capsule networks. Our work differs significantly in studying learning from generic videos instead of from a single visual category.

Learning Objectness from data. Locating objects and separating them against background is a central problem for high-level visual recognition. Prior works on images first generate hierarchical segmentations [36–38] using low-level visual cues such as colors and boundaries [39], and then rank these candidate regions according to a certain criterion [40, 41]. These approaches are able to separate individual object instances, however, with a cost of over-redundant number of proposals. Video-base approaches bring more promises as slow motion analysis [42] is found to be critical for high-level visual development in newborn animals [43, 44]. A simple linear model [45, 46] is shown to factorize a sequence of images into layers of foreground objects and background, assuming motion independence among the objects and background. Later works show application of the layered representations for better optical flow estimates [47, 24, 48] and view-interpolation and time retargeting problems [49–52]. Our model follows the idea of layered representations, and it demonstrates the emergence of objectness from raw video data using self-supervised learning.

Image representation learning from videos. Motion contains rich cues about object location, shape, and part hierarchy. Motion segmentation has been used as a self-supervision signal for learning image-level object representations [53]. Motion propagation [54] predicts a dense optical flow field from sparse optical flow vectors, conditioned on an RGB image. Our work also produces an image representation from unlabeled videos. Unlike prior works, our image representation is a by-product of our full framework for video understanding.

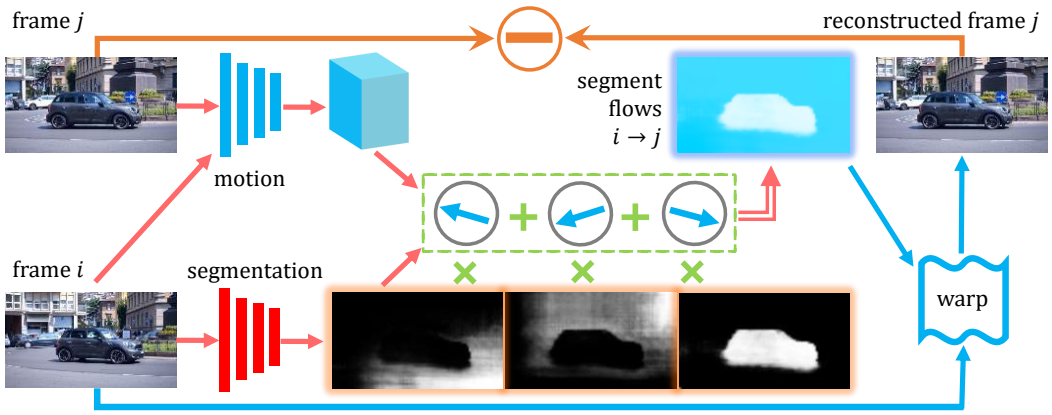

Figure 2: We learn a single-image segmentation network and a dual-frame motion network with an unsupervised image reconstruction loss. We sample two frames, i and j, from a video. Frame i goes through the segmentation network and outputs a set of masks, whereas frames i and j go through the motion network and output a feature map. The feature is pooled per mask and a flow is predicted. All the segments and their flows are combined into a segment flow representation from frame $i \to j$, which are used to warp frame i into j, and compared against frame j to train the two networks.

3 Segmentation by Appearance-Motion Decomposition

The goal of this work is to learn a zero-shot model to detect and segment objects by merely exposing it to unlabeled videos. We are only interested in detecting the objectness in this paper instead of further categorizing the objects into specific classes. We assume that a single moving object appears in one video. When multiple moving objects occur, the model needs to group these objects as one.

We take a learning-based approach to this problem. During training, we are given a collection of unlabeled videos for self-supervised learning. The pretrained model should be directly applicable for inference on a novel image or a video to produce the object segmentation masks. The overall pipeline for training our model is illustrated in Figure 2. Our approach appearance-motion decomposition (AMD) takes a pair of RGB frames X_i and X_j sampled from a video for learning. The model consists of an appearance pathway $f_A(X_i)$ and a motion pathway $f_M(X_i, X_j)$. The two pathways jointly construct a segment flow representation F, which is used to warp frame X_i into X_j. The overall model is self-supervised by the reconstruction objective on the frame X_j. In the following, we describe the details for each module in our model.

3.1 Appearance Pathway for Segmentation

The appearance pathway is a fully convolutional neural network for segmenting a static RGB image into separate regions. Formally, given the image $X_i \in \mathbb{R}^{3\times h\times w}$, it is segmented into c regions by

$$S = f_A(X_i) \in \mathbb{R}^{c\times h\times w}. \tag{1}$$

In practice, the mask S is a soft probability distribution normalized across c channels. c is an important hyper-parameter of our approach. A large c may lead to over-segmentation, and a small c may not locate the object. Empirically, we use a default value of $c = 5$, and this is examined later in an ablation study.

We note that our segmentation network is designed to operate on static images and thus the network can be transferred to downstream image-based vision tasks. In Section 4.2, we demonstrate that the pretrained segmentation network can be used to detect salient objects in a zero-shot fashion. Fine-tuning the appearance pathway on a labeled dataset is examined in Section 4.4.

3.2 Motion Pathway for Correspondence

The purpose of the motion pathway is to extract pixel-wise motion features between a pair of images in order to predict the region flow vector detailed in the next subsection. We follow the network architecture of PWC-Net [20] for predicting dense optical flow, where the feature for each pixel

describes the perceptual similarity to its spatial neighbors in the other frame. Formally, given input frames X_i and X_j, the network extracts features V by

$$V = f_M(X_i, X_j) \in \mathbb{R}^{d_v \times h \times w} \tag{2}$$

where d_v is the dimension of motion features.

3.3 Segment Flow Representation

Given the decoupled appearance pathway and motion pathway, the segment and its motion can be binded for view synthesis. Concretely, we pool the pixel-wise motion features within each segmentation mask to obtain the mask motion feature as a single vector,

$$V_m = \frac{\sum(V \odot S_m)}{\sum S_m} \in \mathbb{R}^{d_v}, \quad m = 1, ..., c \tag{3}$$

where the summation operation is taken across the spatial coordinates, and m is used to index the segmentation masks. The optical flow vector for each segmentation mask is read out from the motion feature by

$$F_m = g(V_m) \in \mathbb{R}^2, \quad m = 1, ..., c \tag{4}$$

where the head network $g(\cdot)$ is chosen as a two-layered multilayer perceptron (MLP).

So far, we decompose a pair of images $X_i \rightarrow X_j$ into a set of segmentation masks S_m and their associated flow vectors F_m. This decomposition is based on the assumption that pixels within a mask share the same motion, a condition that simplifies optical flow estimation. This assumption may not hold for articulated objects and inhomogeneous motion. However, it becomes less problematic when all views in a video are taken for optimization, with the appearance pathway able to aggregate a smoothly moving region into a meaningful segment.

We reconstruct a novel flow representation for the full image by composing the layers of segments with their motion vectors,

$$F = \sum_m F_m \odot S_m, \quad m = 1, ..., c, \tag{5}$$

where $\odot$ denotes the outer product. Since the flow representation F is segment-based, we refer to it as *segment flow*. This decoupled representation allows each component to cross-supervise each other. Given an optical flow offset, the segmentation network could be supervised to find pixels that share this offset. Given a segmentation mask, the correspondence network could be supervised to find the flow offset for this mask.

This approach for supervising object segmentation using motion information is fundamentally different from motion segmentation methods. Our segmentation mask is predicted from a static appearance model that does not require dense and accurate flow for supervision. It utilizes flow at the region level, which can be approximated from sparse and noisy pixel-level estimates.

3.4 Reconstruction Objective

With the segment flow offset map, we are able to warp frame X_i to X_j by

$$\hat{X}_j(p) = X_i(p + F(p)), \tag{6}$$

where p is a spatial location index. The ground-truth frame X_j provides supervision for reconstructed frame $\hat{X}_j$ through the following objective,

$$\mathcal{L} = D(X_j, \hat{X}_j), \tag{7}$$

where D is a metric defining distance between two images. Among the numerous choices for D, such as photometric losses [55], deep-feature-based losses [56][57], and contrastive losses [58], we adopt the pixel-wise photometric loss of SSIM [55] in this work for simplicity.

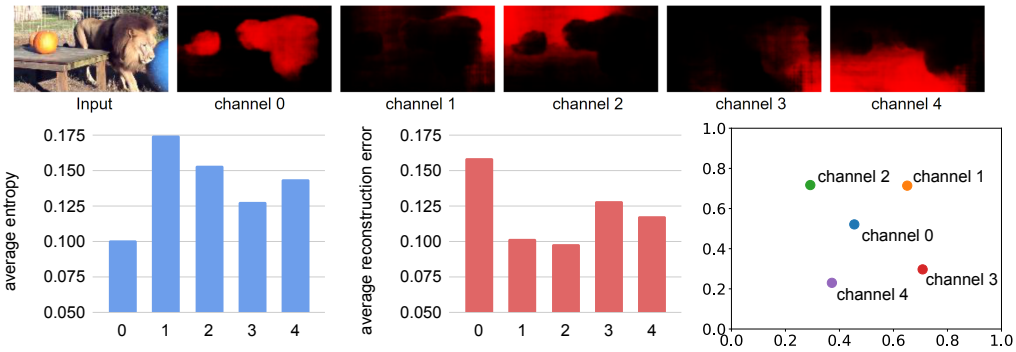

Figure 3: The segment channel which emerges objectness (in this case channel 0) has the least prediction uncertainty, the largest reconstruction error, and mostly lie in the image center. The bottleneck of piece-wise constant motion assumption and limited number of segments encourage the emergence of objectness. Note that the index of the channel which emerges objectness is randomized for different training runs.

3.5 The Emergence of Objectness

The appearance branch of our model segments the input image into c regions, and the region corresponding to the object instead of the background needs to be determined. We have empirically observed that the primary objects all group in a particular mask channel across the training videos. To reveal further understanding about this phenomenon, we calculate three statistics for each channel from 17,500 sampled frames across training videos: per-pixel average of prediction uncertainty measured by entropy, per-pixel average of reconstruction error, and per-pixel average of location. In Figure 3, we find that the channel which corresponds to the objects has the least prediction uncertainty, and the largest reconstruction training error, and mostly lie in the image center. This is consistent to common sense understanding on videos, where objects of focus are put in the center, and piece-wise constant motion assumption holds better for the background. The prediction uncertainty indicates that pixel motion on the object is independent to other regions, while pixel motion on the background may be interpolated from multiple background channels.

3.6 Limitations

The model described is limited to segment foreground objects from the background, without the ability to further differentiate the detected objects into various classes, or separate distinctive object instances from the same class. Also, it is not guaranteed to segment all objects appeared in a visual scene, and sometimes it can only segment part of the object instead of the full shape. The model performance heavily depends on the properties of training videos, the coverage of object categories, the object motion as well as camera motion. Though we find that the model may generalize across datasets in a zero-shot fashion, it is not guaranteed that the model could generalize to unseen objects and unseen scenes during testing time.

4 Experiments

We demonstrate that AMD model can be transferred to three downstream applications. First, the appearance pathway is directly applied on static images for salient object detection in a zero-shot fashion. Second, both the appearance and the motion pathway are transferred to video object segmentation in novel videos with zero human labels. Third, we fine-tune the appearance pathway on labeled data for semantic segmentation.

4.1 Training and Implementation Details

AMD is pretrained on the large object-centric video dataset Youtube-VOS [59]. The training split for Youtube-VOS contains about 4,000 videos covering 94 categories of objects. The total duration for the dataset is 334 minutes. We train the model on the data with a sampling rate 24 frames per second, without using the original segmentation labels.

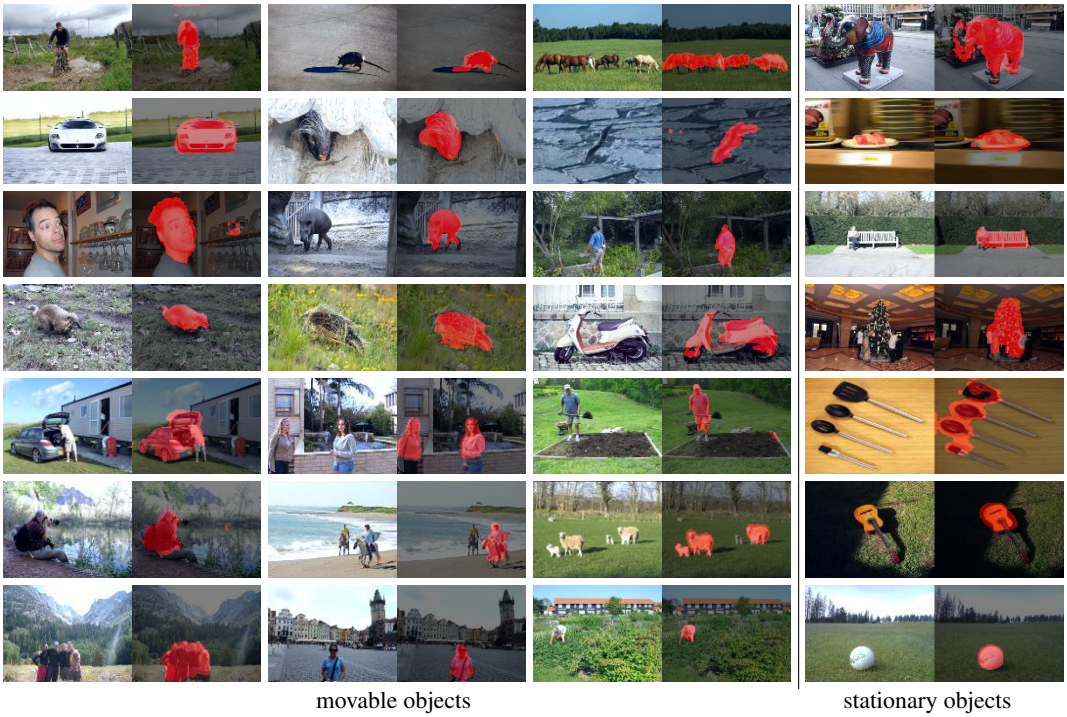

Figure 4: Qualitative salient object detection results. We directly transfer our pretrained segmentation network to novel images on the DUTS dataset without any finetuning. Surprisingly, we find that the model pretrained on videos to segment moving objects can generalize to detect stationary unmovable objects in a static image, e.g. the statue, the plate, the bench and the tree in the last column.

We train all model parameters from scratch without external pretraining. For the segmentation network, we use ResNet50 [60] as our backbone followed by a fully convolutional head containing two convolutional blocks. For the motion network, we adopt PWC-Net [20] architecture because of its effectiveness in estimating optical flows. We resize the short edge of the input image to 400 pixels, and random crop a square image of size 384×384 with random horizontal flipping augmentation. No other augmentations is engineered. We adopt the symmetric loss that considers either frame as the target frame and sums the two reconstruction errors. For training the overall model, we use the Adam optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-6}. We train AMD on eight V100 GPUs, with each processing two pairs of sampled adjacent frames. The network is optimized for 400K iterations.

4.2 Zero-Shot Saliency Detection

Once pretrained, AMD's appearance pathway can be directly transferred to object segmentation in novel stationary images without any downstream fine-tuning. To evaluate the quality of the segmentation, we benchmark the results on salient object detection benchmark.

The salient object detection performance is measured on the **DUTS [61]** benchmark, which contains 5,019 test images with pixel-level ground truth annotations. We follow two widely used metrics in this area: the F_β score and the per-pixel mean squared errors (MAE). F_β is defined as the weighted harmonic mean of the precision (P) and recall (R) scores: $F_\beta = \frac{(1+\beta^2)P \times R}{\beta^2 P + R}$, with $\beta^2 = 0.3$. MAE is simply the per-pixel averaged error of the soft prediction scores.

Experimental results. We compare our saliency estimation results against several traditional methods based on low-level cues. Useful low-level cues and priors include background priors [62], objectness [63, 64], and color contrast [65]. As shown in Table 1, our method achieves an F_β score 60.2 and an MAE score of 0.13, outperforming all traditional approaches by a notable margin. We note that AMD is not designed specifically for this task nor for this particular dataset, and its strong performance demonstrates the generalization ability of the model.

Table 1: Salient object detection performance on the DUTS dataset. Our model outperforms traditional low-level methods by notable margins.

Model	F_β	MAE
RBD[62]	51.0	0.20
HS[69]	52.1	0.23
MC[63]	52.9	0.19
DSR[70]	55.8	0.14
DRFI[64]	55.2	0.15
AMD	**60.2**	**0.13**

Table 2: Transfer performance for semantic segmentation on VOC2012. Our method outperforms TimeCycle and compares favorably with contrastive methods.

Model	Data	Aug.	mIoU
Scratch	–	–	48.0
TimeCyle[71]	VLOG	light	52.8
MoCo-v2[2]	YTB	light	61.5
AMD	YTB	light	**62.0**
MoCo-v2[2]	YTB	heavy	**62.8**
AMD	YTB	heavy	62.1
MoCo-v2[2]	IMN	heavy	**72.4**

In related work on unsupervised learning of saliency detection [66–68], the priors of traditional low-level methods are ensembled. Though they do not use saliency annotations, their models are pretrained on ImageNet classification and even semantic segmentation with pixel-level annotations. These methods are thus not fully unsupervised, so they are omitted from the comparisons.

In Figure 4, we show some qualitative results on salient object detection. Surprisingly, we find that our model pretrained on videos to segment moving objects not only detects movable objects in images, but also generalizes to detect stationary unmovable objects, such as statues, benches, trees and plates shown in the last column. This suggests that our model learns a generic objectness prior from the unlabeled videos. To further quantify this, we manually label the objects from the DUTS dataset into movable objects and stationary objects. The F1 score of movable objects and stationery objects are 63.0 and 57.9 respectively, without significant performance gap. We hypothesize that our model may learn objectness from the camera motion as well. Camera motion may cause the object and the background at various depth to have different observed projective 2D optical flow even though the objects are static.

4.3 Zero-shot Video Object Segmentation

We transfer the pretrained AMD model to object segmentation on novel videos. Since the segmentation prediction from our model is based on static images, inference on images sequentially in a video essentially estimates objectness. In order to exploit motion information, we use a test-time adaptation approach. Concretely, given a novel video, we optimize the training objective in Eq. 7 on pairs of frames sampled from the novel testing video. The adaptation takes 100 iterations per video.

We evaluate zero-shot video object segmentation on three testing datasets. **DAVIS 2016 [72]** contains 20 validation videos with 1,376 annotated frames. **SegTrackv2 [73]** contains 14 videos with 976 annotated frames. Following prior works, we combine multiple foreground objects in the annotation into a single object for evaluation. **FBMS59 [27]** contains 59 videos with 720 annotated frames. The dataset is relatively challenging because the object may be static for a period of time. We pre-process the ground truth labels following prior work [18]. For evaluation, we report the Jaccard score, which is equivalent to the intersection over union (IoU) between the prediction and the ground truth segmentation.

Experimental results. We consider baseline methods claiming to be unsupervised for the full pipeline, including traditional non-learning-based approaches [74, 14, 16, 28, 15] and recent self-supervised learning methods [18, 19]. In Table 3, we summarize the results for all the methods on the three datasets. Among these methods, NLC [14] actually relies on an edge model trained with human-annotated edge boundaries, and ARP [15] depends on a segmentation model trained on a human-annotated saliency dataset. We thus gray their entries in the table. For all the traditional methods, since their original papers do not report results on most of these benchmarks, we simply provide the performance values reported in the CIS paper [18].

We evaluate the performance for AMD with and without test-time adaptation. No adaptation boils down to per-image saliency estimation using only the appearance pathway. Adaptation transfers both appearance and motion pathways. On DAVIS 2016, our method achieves a Jaccard score of 57.8%, surpassing all traditional unsupervised models. For CIS [18], their best performing model uses a significant amount of post-processing, including model ensembing, multi-crop, temporal smoothing and spatial smoothing. We thus refer to their performance obtained from a single model without post-

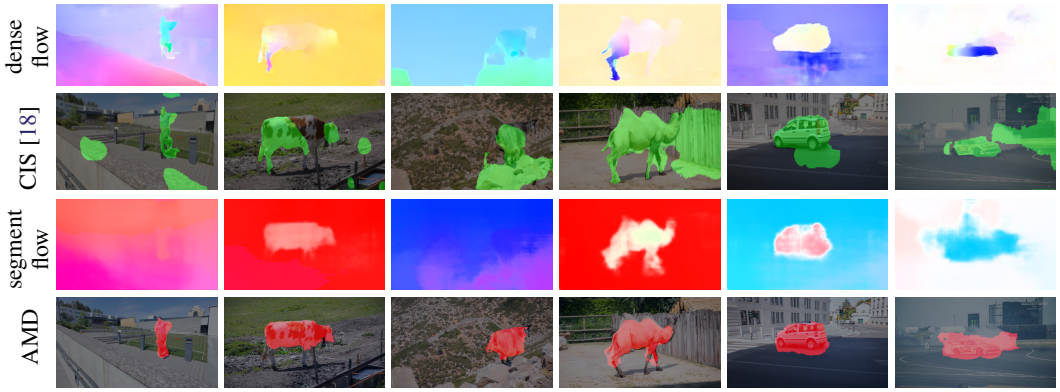

Figure 5: Qualitative comparisons to motion segmentation based method CIS [18] with its input dense flow and our segmentation results with segment flow representation. CIS is prone to noise, articulated motion, and camera motion in the dense flow estimations. By decomposing appearance from motion, our model AMD suffers less from these vulnerabilities of dense optical flow. This leads to results much better and more robust than the motion segmentation based approach.

Table 3: Performance evaluations for unsupervised video object segmentation on DAVIS 2016, SegTrackv2 and FBMS59 datasets. The numbers are measured in terms of Jaccard score. The table is split into traditional non-learning-based and recent self-supervised learning methods. The model results which rely on other kinds of human supervisions (Sup.) are grayed. Dependence for pretrained dense flow method is also listed for each model. MG's results on SegTrackv2 and FMBS59 using ARFlow are reproduced by ours and marked with $*$. We evaluate AMD with appearance pathway only and with both pathways for test time adaptation. AMD performs favorably to the baseline CIS on DAVIS 2016, while showing large gains on the other two benchmarks.

	Model	e2e	Sup.	Flow	DAVIS 2016	SegTrackv2	FBMS59
traditional	SAGE[74]	✗	✗	LDOF[75]	42.6	57.6	61.2
	NLC[14]	✗	edge	SIFTFlow[76]	55.1	67.2	51.5
	CUT[28]	✗	✗	LDOF[75]	55.2	54.3	57.2
	FTS[16]	✗	✗	LDOF[77]	55.8	47.8	47.7
	ARP[15]	✗	saliency	CPMFlow[78]	76.2	57.2	59.8
learning	CIS[18]	✗	✗	PWC[20]	**59.2**	45.6	36.8
	MG[19]	✗	✗	ARFlow[6]	53.2	37.8*	**50.4***
	AMD (per-img)	✓	✗	✗	45.7	28.7	42.9
	AMD (per-vid)	✓	✗	✗	57.8	**57.0**	47.5

processing. Our model is slightly worse than CIS on DAVIS, by 1.4%. However, on SegTrackv2 and FBMS59, our method outperforms CIS by large margins of 11.4% and 10.7%, respectively. Motion grouping [19] is a work concurrent with ours. It is essentially a motion segmentation approach which relies on an off-the-shelf pre-computed dense optical flow model. Motion grouping performs worse than our method on DAVIS2016 and SegTrackv2 when a low-performance unsupervised optical flow model ARFlow is used [6]. With a state-of-the-art supervised optical flow model [21] that is trained on ground truth flow, their performance improves significantly. Among all the discussed methods, ours is the first end-to-end self-supervised learning approach which does not require a pretrained optical flow model.

In Figure 5, we show qualitative comparisons to the baseline CIS [18]. We display dense flow from pretrained PWC-Net, the CIS results, our segment flows, and our segmentation results. For most of these examples, our segment flow only coarsely reflects the true pixel-level optical flow. However, our segmentation results are significantly better and less noisy, as our model is relatively insensitive to optical flow quality. In the first and the third examples, our model produces high-quality object segmentations even though the motion cue for the objects is very subtle.

#segments	DAVIS (J)
$c = 5$	**57.8**
$c = 6$	45.3
$c = 8$	41.0

Figure 6: Ablation study on different number of segments. Two examples with segmentation masks and segment flows are shown. The object region is split over multiple masks when c becomes large. The over-segmentation decreases the performance of video object segmentation on DAVIS2016.

4.4 Semantic Segmentation

Given that our pretrained segmentation network can produce meaningful generic object segmentations, we further examine its semantic modeling ability on semantic segmentation. We conduct this experiment on the **Pascal VOC 2012 [79]** dataset. The dataset contains 20 object categories with 10,582 training images and 1,449 validation images. Given a pretrained model, we finetune the model on the training set and evaluate the performance on the validation set. The finetuning takes 40,000 iterations with a batch size of 16 and an initial learning rate of 0.01. The learning rate undergoes polynomial decay with a power parameter of 0.9.

Experimental results. We compare the pretrained model to a image-based contrastive model, MoCo-v2 [2] and a self-supervised video pretraining model, TimeCycle [71]. TimeCycle [71] is pretrained on the VLOG dataset, which is larger than our Youtube-VOS dataset. For MoCo-v2, we also pretrain the contrastive model on the Youtube-VOS dataset, to ablate the role of pretraining datasets. Since the base version of our method does not utilize heavy augmentations as in contrastive models, we also study the effects of data augmentations. The results are reported in Table 2. Our method outperforms the video pretraining approach TimeCylce significantly by 9.2%. Compared with MoCo-v2, when light augmentation (resizing, cropping) is used, our model slightly outperforms MoCo-v2 by 0.5%. However, when heavy data augmentation (color jitter, grayscale, blurring) is applied, our method underperforms MoCo-v2 by 0.7%. This is possibly because our model is non-contrastive in nature, and thus unable to take advantage of information effectively in augmentations. MoCo-v2 performs much stronger when pretrained on ImageNet, possibly because the semantic distribution of ImageNet is well aligned with VOC2012. Overall, our model outperforms a prior self-supervised video model TimeCycle and compares favorably with contrastive model MoCo-v2 under the same data.

4.5 Ablation Study

The variable c, the number of segmentation channels, is an important hyper-parameter of our model. We vary the value of c $(5, 6, 8)$ for pretraining the model, and examine its transfer performance for video object segmentation on DAVIS2016. In Figure 6, we visualize the model predictions under different number of segments. We observe that a large c tends to lead to over-segmentation, and a small c tends to lead to large regions. The car and the swan is split into multiple regions even if the motion for separated regions are very close. The model trained with $c = 5$ segments a full object, while the model trained with $c = 8$ separates the object into parts. When pretraining the model with $c < 4$, the training becomes unstable. Quantitatively, the video object segmentation performance on DAVIS2016 decreases as we increase the number of segments.

5 Summary

In this paper, we show that objectiveness could emerge from a computational model by exposing it to unlabeled videos. We present a novel model which decouples the appearance pathway and the motion pathway, and later binds them into a joint segment flow representation. As opposed to prior works that rely heavily on accurate dense optical flows for predicting object segmentation, our method learns only from raw pixel observations. The motion representation in our model is a lot weaker, however the object segmentation is more robust. The proposed model AMD is the first end-to-end learning approach for zero-shot object segmentation without using any pretrained modules. Its performance is validated on a number of image and video object segmentation benchmarks.

References

[1] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

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

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

[5] 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.

[6] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6489–6498, 2020.

[7] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[8] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[9] Max Wertheimer. Untersuchungen zur lehre von der gestalt. ii. *Psychologische forschung*, 4(1):301–350, 1923.

[10] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019.

[11] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6949–6958, 2020.

[12] 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.

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

[14] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, volume 2, page 8, 2014.

[15] Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation and reduction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7417–7425. IEEE, 2017.

[16] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE international conference on computer vision*, pages 1777–1784, 2013.

[17] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3395–3402, 2015.

[18] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 879–888, 2019.

[19] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. *arXiv preprint arXiv:2104.07658*, 2021.

[20] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[21] 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.

[22] 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.

[23] M Pawan Kumar, Philip HS Torr, and Andrew Zisserman. Learning layered motion segmentations of video. *International Journal of Computer Vision*, 76(3):301–319, 2008.

[24] Deqing Sun, Erik B Sudderth, and Michael J Black. Layered segmentation and optical flow estimation over time. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1768–1775. IEEE, 2012.

[25] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3386–3394, 2017.

[26] Pavel Tokmakov, Cordelia Schmid, and Karteek Alahari. Learning to segment moving objects. *International Journal of Computer Vision*, 127(3):282–301, 2019.

[27] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1187–1200, 2013.

[28] Margret Keuper, Bjoern Andres, and Thomas Brox. Motion trajectory segmentation via minimum cost multicuts. In *Proceedings of the IEEE international conference on computer vision*, pages 3271–3279, 2015.

[29] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7334–7344, 2019.

[30] 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.

[31] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 869–878, 2019.

[32] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. *arXiv preprint arXiv:1806.07823*, 2018.

[33] Zhenjia Xu, Zhijian Liu, Chen Sun, Kevin Murphy, William T Freeman, Joshua B Tenenbaum, and Jiajun Wu. Unsupervised discovery of parts, structure, and dynamics. *arXiv preprint arXiv:1903.05136*, 2019.

[34] Aliaksandr Siarohin, Subhankar Roy, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Motion-supervised co-part segmentation. *arXiv preprint arXiv:2004.03234*, 2020.

[35] Sara Sabour, Andrea Tagliasacchi, Soroosh Yazdani, Geoffrey E Hinton, and David J Fleet. Unsupervised part representation by flow capsules. *arXiv preprint arXiv:2011.13920*, 2020.

[36] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering occlusion boundaries from an image. *International Journal of Computer Vision*, 91(3):328–346, 2011.

[37] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

[38] Eitan Sharon, Meirav Galun, Dahlia Sharon, Ronen Basri, and Achi Brandt. Hierarchy and adaptivity in segmenting visual scenes. *Nature*, 442(7104):810–813, 2006.

[39] Ian Endres and Derek Hoiem. Category-independent object proposals with diverse ranking. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):222–234, 2013.

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

[41] Sudheendra Vijayanarasimhan and Kristen Grauman. Efficient region search for object detection. In *CVPR 2011*, pages 1401–1408. IEEE, 2011.

[42] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.

[43] Justin N Wood and Samantha MW Wood. The development of newborn object recognition in fast and slow visual worlds. *Proceedings of the Royal Society B: Biological Sciences*, 283(1829):20160166, 2016.

[44] Justin N Wood. A smoothness constraint on the development of object recognition. *Cognition*, 153:140–145, 2016.

[45] John YA Wang and Edward H Adelson. Layered representation for motion analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 361–366. IEEE, 1993.

[46] John YA Wang and Edward H Adelson. Representing moving images with layers. *IEEE transactions on image processing*, 3(5):625–638, 1994.

[47] Deqing Sun, Jonas Wulff, Erik B Sudderth, Hanspeter Pfister, and Michael J Black. A fully-connected layered model of foreground and background flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2451–2458, 2013.

[48] Deqing Sun, Erik Sudderth, and Michael Black. Layered image motion with explicit occlusions, temporal consistency, and depth ordering. *Advances in Neural Information Processing Systems*, 23:2226–2234, 2010.

[49] Gabriel J Brostow and Irfan A Essa. Motion based decompositing of video. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 8–13. IEEE, 1999.

[50] Jean-Baptiste Alayrac, Joao Carreira, and Andrew Zisserman. The visual centrifuge: Model-free layered video representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2457–2466, 2019.

[51] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM transactions on graphics (TOG)*, 23(3):600–608, 2004.

[52] Erika Lu, Forrester Cole, Tali Dekel, Weidi Xie, Andrew Zisserman, David Salesin, William T Freeman, and Michael Rubinstein. Layered neural rendering for retiming people in video. *arXiv preprint arXiv:2009.07833*, 2020.

[53] 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.

[54] Xiaohang Zhan, Xingang Pan, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised learning via conditional motion propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1881–1889, 2019.

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[56] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[58] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020.

[59] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

[60] 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.

[61] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017.

[62] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2814–2821, 2014.

[63] Bowen Jiang, Lihe Zhang, Huchuan Lu, Chuan Yang, and Ming-Hsuan Yang. Saliency detection via absorbing markov chain. In *Proceedings of the IEEE international conference on computer vision*, pages 1665–1672, 2013.

[64] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2083–2090, 2013.

[65] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014.

[66] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9029–9038, 2018.

[67] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, and Yizhou Yu. Multi-source weak supervision for saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6074–6083, 2019.

[68] Duc Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction with self-supervision. *arXiv preprint arXiv:1909.13055*, 2019.

[69] Wenbin Zou and Nikos Komodakis. Harf: Hierarchy-associated rich features for salient object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 406–414, 2015.

[70] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE international conference on computer vision*, pages 2976–2983, 2013.

[71] 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.

[72] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[73] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2192–2199, 2013.

[74] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):20–33, 2017.

[75] Thomas Brox and Jitendra Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):500–513, 2010.

[76] Ce Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009.

[77] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European conference on computer vision*, pages 438–451. Springer, 2010.

[78] Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *CVPR*, 2016.

[79] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

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] All data used in the paper is public; our code will be published.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1 and Section 4.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Unfortunately, we did not provide the error bars due to resource constraints. We will make this up in the future.

(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] See Section 4.1.

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] We cited all video and image datasets used in this paper.

   (b) Did you mention the license of the assets? [No]

   (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? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]

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? [N/A]

   (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? [N/A]