Zero-Shot Object Segmentation by Appearance-Motion Decomposition

Anonymous Author(s)
Affiliation
Address
email

Abstract

Humans can easily detect and segment moving objects simply by observing how they move, even without knowledge of object semantics. Inspired by this, we develop a zero-shot unsupervised approach for video object segmentation. The model comprises two visual pathways: an appearance pathway that segments individual RGB images into coherent object regions, and a motion pathway that predicts the flow vector for each region between consecutive video frames. The two pathways jointly reconstruct a new representation called *segment flow*. This decoupled representation of appearance and motion is trained in a self-supervised manner to reconstruct one frame from another.

When pretrained on an unlabeled video corpus, the model can be useful for a variety of applications, including 1) primary object segmentation from a single image in a zero-shot fashion; 2) moving object segmentation from a video with unsupervised test-time adaptation; 3) image semantic segmentation by supervised fine-tuning on a labeled image dataset. We demonstrate encouraging experimental results on all of these tasks using pretrained models.

1 Introduction

Humans learn to perceive the world through a continuous stream of visual sensory input. A dynamic sequence of observations provides information about what is moving in the scene and how it moves. Such motion patterns not only reveal the boundary segment of an object, but also indicates hierarchical part structures and object class. Thus, studying motion patterns can be of great value for uncovering low-level, mid-level and high-level object representations.

Segmenting moving objects in videos is a longstanding research problem in computer vision. A well-known Gestalt principle from 1923 [1] states the rule of *common fate* where pixels that move together belong together. Classical methods for motion segmentation [2–4] implement this idea by separating distinctive motion regions from the background based on two-frame optical flow. Recent self-supervised methods [5, 6] follow a similar intuition that motions of the foreground and background are statistically independent of each other. These motion-based approaches to video object segmentation are fully subject to the per-pixel optical flow representation.

The requirement of dense and accurate optical flow may be problematic in the following aspects as shown in Figure 1. First, the estimated optical flow may not be smooth over time, thus leading to temporally inconsistent predictions. Second, optical flow is vulnerable to articulated objects with inhomogeneous motion [7]. Third, estimating optical flow itself is an ill-posed problem that involves reasoning about 3D geometry from 2D images.

In this paper, we raise the question of whether dense optical flow is necessary for segmenting moving objects in videos. In answering this, we turn our attention to modeling appearance on

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

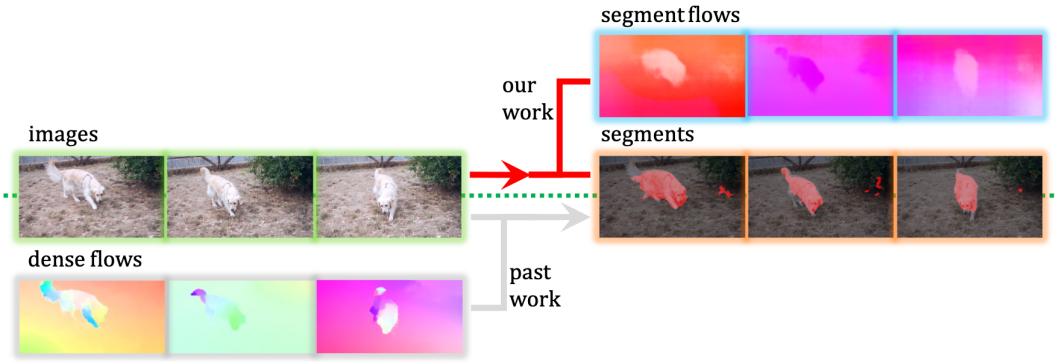


Figure 1: Our zero-shot object segmentation is learned from an unsupervised factorization of images into segments and their motions, whereas past work segments objects based on dense pixel-wise optical flows, which are brittle in the presence of noise, articulated movement, and abrupt motion.

RGB representations. Appearance modeling in a video is beneficial for segmenting objects in two aspects. First, appearance variations across time tend to be much smoother than motion variations. Second, appearance provides rich cues (e.g. texture, color and edges) for perceptual organization, which can alleviate the need for dense pixel correspondence. We also note the related literature on supervised/semi-supervised video object segmentation [8, 9] that do not rely on an explicit optical flow representation.

In this paper, we propose an end-to-end self-supervised learning approach for unsupervised video object segmentation. Our approach adopts two visual pathways: an appearance pathway to model "what is moving" by segmenting individual RGB images into coherent regions, and a motion pathway to model "how it moves" by predicting a single optical flow vector for each region assuming common fate. In this way, object appearance and motion are decoupled, such that the appearance model for predicting segmentation would benefit from smoothly varying visual signals. By conditioning on each region, the motion pathway is also tasked to solve a much simpler task than dense optical flow. The segmentation masks and flow vectors jointly reconstruct a new representation called *segment flow*, which is used to warp the source frame to reconstruct the target frame. The two pathways are jointly trained with a reconstruction loss in a self-supervised fashion.

After unsupervised pretraining, our model has the versatility to be used in a variety of applications. First, the appearance pathway can directly be 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, the overall model can be transferred to novel videos for video object segmentation in a zero-shot fashion, with unsupervised test-time adaptation to the novel videos. 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 end-to-end self-supervised learning framework for unsupervised video object segmentation without any pretrained external modules; 2) segmentation of objects from unlabeled videos without requiring dense and accurate optical flow; 3) a versatile model that can be applied to various image and video segmentation tasks in a zero-shot fashion.

2 Related Works

Intrinsic Decomposition. The observed RGB values of image pixels arise from the mixed interactions of albedo, shading, motion, and shapes, among others. A decomposition of these factors into what is commonly known as intrisic images [10] can be of great value for recognition and segmentation purposes, as invariant representations can be derived from these physical characteristics. Intrinsic image estimation has been formulated in a self-supervised manner [11] by using a reconstruction loss and enforcing various physical constraints on the intrinsic images. Depth, optical flow, and object shapes are other physical properties which can be recovered as intrinsic images from videos [12, 13].

In a similar spirit, our work aims to recover object shapes and motions by decomposing them as independent physical properties from video.

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 [8, 14], or learned through self-supervision such as colorization [15] and cycle-consistency [9]. 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 [16] and ARP [17] 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 [18] calculates the segmentation by obtaining a motion boundary from the optical flow map between frames. SAGE [19] takes into account multiple cues of edges, motion segmentation, and image saliency for video object segmentation. Contextual information separation [5] segments moving objects by exploiting the motion independence between the foreground and the background. A concurrent work based on motion grouping [6] 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 [22] supervision.

Motion Trajectory Segmentation. Moving object segmentation has been shown to be effective when motion is considered over a large time interval [7]. An approach based on trajectory clustering [23] 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.

Layered representations. A simple linear model [24, 25] can factorize a video into layers of foreground objects and background, assuming independence among the objects and background. This layered representation was used to derive better optical flow estimates [26, 4, 27] and also for view-interpolation and time retargeting applications [28–31]. Our work also constructs a layered representation in which object appearance and motion are decomposed. Different from prior works, we show that meaningful segmentations can emerge from this end-to-end framework in a fully unsupervised manner.

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 [32] predicts segmentation by learning to group super-pixels of similar appearance and context through contrastive learning. Later work [33] contrasts holistic mask proposals obtained from traditional bottom-up grouping.

A related line of work focuses on learning part segmentation from images of the same object category, such as humans and faces. SCOPS [34] is a representative method learned in a self-supervised fashion. The general idea follows unsupervised landmark detection [35], where geometric invariance, representation equivariance and perceptual reconstructions are considered. Co-part segmentation [36] is also explored in videos, where motion provides a strong cue for part organization. A motion-supervised approach [37] models part motion between adjacent frames via affine parameters. Recent work [38] implements a similar idea in capsule networks.

Image representation learning using motions. 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 [39]. Motion propagation [40] 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.

3 Segmentation by Appearance-Motion Decomposition

Given a video, our goal is to estimate the segmentation mask for the moving object in all frames without using any labels. In this work, a single moving object is assumed. When multiple moving objects are present, the model needs to segment multiple objects as a single one. Moving object segmentation a challenging task due to camera motion, articulated object motion, and occlusions.

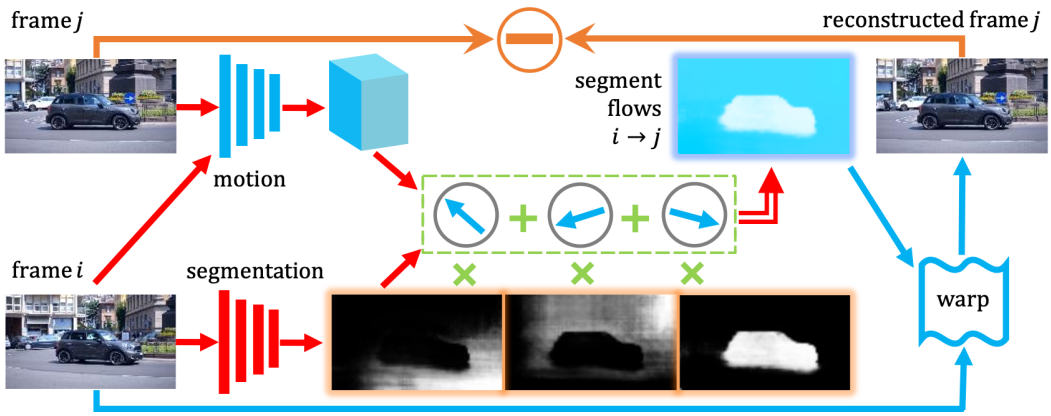


Figure 2: We learn a single-image segmentation network and a dual-frame motion network with an unsupervised image reconstruction loss. We sample two frames, i and j, from a video. Frame i goes through the segmentation network and outputs a set of masks, whereas frames i and j go through the motion network 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 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.

We take a learning-based approach to this problem. During training, we are given a collection of unlabeled videos for unsupervised learning. After training, the pretrained model should be directly applicable for inference on a novel unlabeled video to produce the object segmentation mask. Note that in both training and inference, the model does not exploit any human labels.

The overall pipeline for training our model is illustrated in Figure 2. Our approach 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 layers. Formally, given the image $X_i \in \mathbb{R}^{3\times h\times w}$, it is segmented into c layers 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 on a labeled dataset is examined in Section 4.3.

3.2 Motion Pathway for Correspondence

The purpose of the motion pathway is to predict the optical flow vector for each segmentation mask obtained from the appearance pathway. The network first extracts pixel-wise motion features V from the input frames X_i and X_j,

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

where d_v is the dimension of motion features. We then 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 a linear layer.

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 frames in a video are utilized for optimization, with the appearance pathway able to aggregate a smoothly moving region into a coherent segment.

3.3 Segment Flow Representation

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

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

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 optical flow vector 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 [41], deep-feature-based losses [42][43], and contrastive losses [44], we adopt the pixel-wise photometric loss of SSIM [41] in this work for simplicity.

3.5 Primary Object Segment Selection

Since our model outputs c masks for each image, the layer corresponding to the moving object needs to be determined. We have empirically observed that the moving objects all appear in a particular mask channel across the training videos. This channel can be heuristically identified as the one whose segmentation mask is closest to the image center. The object segment from this layer is used for evaluating zero-shot downstream tasks.

3.6 Implementation Details

For the segmentation network, we use ResNet50 [45] as our backbone followed by an FCN head consisting of two convolutional layers. For the motion network, we adopt PWC-Net [20] because of its effectiveness in estimating optical flows. Both pathways are initialized from scratch without any pretraining. We resize the short edge of the input image to 400 pixels, and random crop a square image of size 384×384. 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 the model on eight V100 GPUs, with each processing two pairs of sampled adjacent frames. The network is optimized for 100K iterations.

4 Experiments

We empirically validate our approach for video object segmentation. Our model is pretrained on the large object-centric video dataset Youtube-VOS [46]. The training split for Youtube-VOS contains about 4000 videos covering 94 categories of objects. The total duration for the dataset is 334 minutes. We train our model on the data with a sampling rate 24 frames per second, without using the original segmentation labels.

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

4.1 Zero-shot Video Object Segmentation

We transfer our pretrained 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 for model transfer. Concretely, given a novel video, we optimize the training objective in Eq. 7 on pairs of frames sampled from the novel video. The adaptation takes 100 iterations per video.

We evaluate zero-shot video object segmentation on three testing datasets. **DAVIS 2016 [47]** contains 20 validation videos with 1376 annotated frames. **SegTrackv2 [48]** contains 14 videos with 976 frames that are annotated. Following prior works, we combine multiple foreground objects in the annotation into a single object for evaluation. **FBMS59 [7]** 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 [5]. 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 [49, 16, 18, 23, 17] and recent self-supervised learning methods [5, 6]. In Table 1, we summarize the results for all the methods on the three datasets. Among these methods, NLC [16] actually relies on an edge model trained with human-annotated edge boundaries, and ARP [17] 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 [5].

On DAVIS 2016, our method achieves a Jaccard score of 58.1%, surpassing all traditional unsupervised models. For CIS [5], their best performing model uses a significant amount of post-processing, including model ensembing, multi-crop, temporal smoothing and spatial smoothing. We therefore refer to their performance obtained from a single model without post-processing. Our model is slightly worse than CIS on DAVIS, by 1.1%. However, on SegTrackv2 and FBMS59, our method outperforms CIS by large margins of 12.1% and 16.3%, respectively. Motion grouping [6] is an unpublished work concurrent with ours. It relies on an off-the-shelf pre-computed dense optical flow model. Motion grouping performs worse than our method when a low-performance unsupervised optical flow model is used [22]. 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 that does not require a pretrained optical flow model.

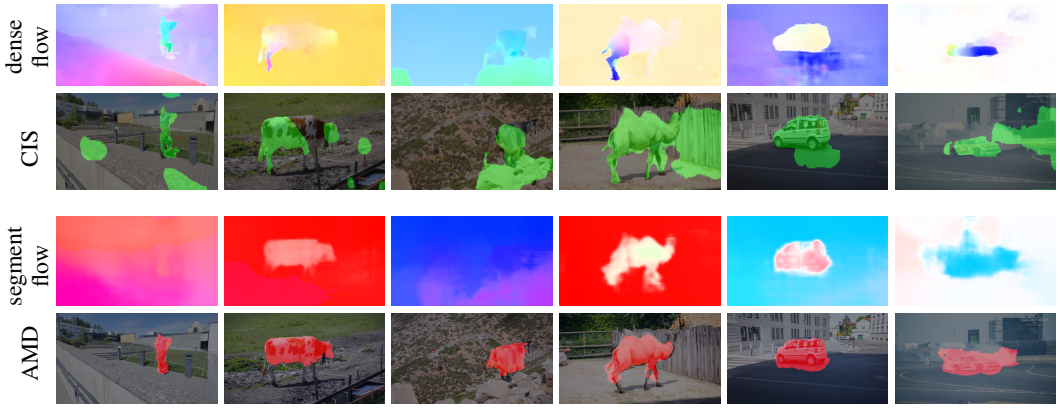


Figure 3: Qualitative comparisons to motion segmentation based method CIS [5]. CIS is prone to noise, articulated motion, camera motion in dense flow estimations. Our segment flow is less accurate than pretrained pixel-wise optical flow. By decomposing appearance from motion, our object segmentation results are much better and robust than the dense flow based approach.

Table 1: Performance comparison for unsupervised video object segmentation in terms of Jaccard score. The table is split into non-learning-based methods and recent self-supervised learning methods. The models which rely on human labels are grayed. Our model is the first end-to-end self-supervised learning approach without relying on pretrained optical flow models. Note that MG [6] is a non-published work currently with ours.

Model	E2E	Sup.	Flow	DAVIS 2016	SegTrackv2	FBMS59
SAGE[49]	✗	✗	LDOF [50]	42.6	57.6	61.2
NLC[16]	✗	edge	SIFTFlow [51]	55.1	67.2	51.5
CUT[23]	✗	✗	LDOF [50]	55.2	54.3	57.2
FTS[18]	✗	✗	LDOF [52]	55.8	47.8	47.7
ARP[17]	✗	saliency	CPMFlow [53]	76.2	57.2	59.8
CIS[5]	✗	✗	PWC [20]	59.2	45.6	36.8
MG[6]	✗	✗	ARFlow [22]	53.2	-	-
MG[6]	✗	✗	RAFT [21]	68.3	58.6	53.1
AMD	✓	✗	✗	58.1	57.7	51.0

In Figure 3, we show qualitative comparisons to the baseline CIS [5]. 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 segment flow reveals little about the moving objects.

4.2 Zero-Shot Saliency Detection

Our pretrained appearance pathway can be directly transferred to segmentation of moving objects in novel stationary images. To evaluate the quality of the segmentations, we benchmark the results on salient object detection.

The salient object detection performance is measured on the **DUTS [54]** 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 [55], objectness [56, 57], and color contrast [58]. As shown in Table 2, our method achieves an F_β score

Figure 4: Qualitative salient object detection results. Surprisingly, we find that the model pretrained on videos to segment moving objects can generalize to detect stationary objects in a static image, e.g. the elephant statue and the plates.

60.2 and an MAE score of 0.13, outperforming all traditional approaches by a large margin. We note that our pretrained model is not designed specifically for this task, and its strong performance demonstrates the generalization ability of the model.

In related work on unsupervised learning of saliency detection [59–61], 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 objects, such as elephant statues and plates. This suggests that our model learns a generic objectness prior from the unlabeled videos.

4.3 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 [62]** 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 two image-based contrastive models, MoCo-v2 [63] and SimSiam [64]. For fair comparison, we pretrain the two baseline representations on the same Youtube-VOS dataset. We also prolong the pretraining of our model to 400K iterations to match the baselines. Since the base version of our method does not utilize heavy augmentations as in contrastive models, we also study the effects of data augmentations for our method and the contrastive baselines. The results are reported in Table 3. When heavy augmentation is not used, our model slightly outperforms MoCo-v2 by 0.5% and SimSiam by 4.3%. However, when heavy data augmentation 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 in augmentations. Overall in this experiment, our model compares favorably with contrastive models. Formulating our model in a contrastive manner is an interesting direction for future study.

4.4 Ablation Study

The variable c, the number of segmentation channels, is an important hyper-parameter of our model. Large c tends to lead to over-segmentation, and small c tends to lead to very large regions. We conduct ablations on this parameter in this section.

Table 2: Salient object detection performance on the DUTS dataset. Our model outperforms traditional low-level methods by large margins.

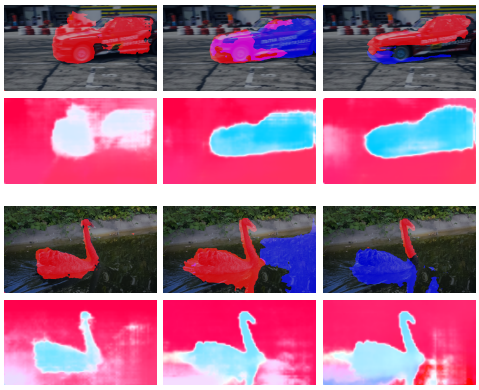
Model	F_β	MAE
RBD[55]	51.0	0.20
HS[65]	52.1	0.23
MC[56]	52.9	0.19
DSR[66]	55.8	0.14
DRFI[57]	55.2	0.15
AMD	**60.2**	**0.13**

Table 3: Transfer performance for semantic segmentation on VOC2012. Our method compares favorably with contrastive methods.

Model	Aug.	mIoU
Scratch	–	48.0
SimSiam[64]	✗	57.7
MoCo-v2[63]	✗	61.5
AMD	✗	**62.0**
SimSiam[64]	✓	59.4
MoCo-v2[63]	✓	**62.8**
AMD	✓	62.1

Table 4: Ablation study on video object segmentation and semantic segmentation on the number of segments c in our model. The results show that a smaller number of segments perform better for video object segmentation and a larger value performs better for semantic segmentation.

#segments	DAVIS (J)	VOC (mIoU)
$c = 5$	**58.1**	59.1
$c = 6$	45.3	59.4
$c = 8$	41.0	**60.4**



$c = 5$ $c = 6$ $c = 8$

Figure 5: 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.

We vary the value of c $(5, 6, 8)$ for pretraining the model, and examine its transfer performance for zero-shot video object segmentation and fine-tuned performance for semantic segmentation. In Table 4, the video object segmentation performance on DAVIS decreases as we increase the number of segments. In Figure 5, we visualize the model predictions under different number of segments. 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.

For semantic segmentation on VOC2012, increasing the value of c increases the transfer performance. We hypothesize that it is because the model trained with $c = 8$ learns a better representation for object semantics by distinguishing object parts. For the model trained with $c = 5$, it is less able to recognize object categories, as all objects occur in the same mask layer. Investigating self-supervised models capable of both segmenting the full envelope of an object and recognizing its semantics is an open direction for future study.

5 Summary

We propose a novel method based on appearance-motion decomposition for zero-shot object segmentation. Our model is the first end-to-end learning-based formulation of zero-shot video object segmentation without using any pretrained modules. As opposed to traditional works that rely heavily on dense optical flows for predicting object segmentations, our method learns only from weak motion supervision. As a result, the flow representation in our model is less accurate, but the object segmentation is more robust. The results on salient object detection suggests that our model learns a generic objectness prior that generalizes across object categories.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2(3-26):2, 1978.

[11] Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Joshua B Tenenbaum. Self-supervised intrinsic image decomposition. *arXiv preprint arXiv:1711.03678*, 2017.

[12] Naejin Kong, Peter V Gehler, and Michael J Black. Intrinsic video. In *European Conference on Computer Vision*, pages 360–375. Springer, 2014.

[13] Naejin Kong and Michael J Black. Intrinsic depth: Improving depth transfer with intrinsic images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3514–3522, 2015.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[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] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

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

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

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

[65] 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.

[66] 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.

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 3.6 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 3.6.
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]