Time-Supervised Primary Object Segmentation

Yanchao Yang* Brian Lai* and Stefano Soatto
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095

Abstract

We describe an unsupervised method to detect and segment portions of live scenes that, at some point in time, are seen moving as a coherent whole, which we refer to as primary objects. Our method first segments motions by minimizing the mutual information between partitions of the image domain, which bootstraps a static object detection model that takes a single image as input. The two models are mutually reinforced within a feedback loop, enabling extrapolation to previously unseen classes of objects. Our method requires video for training, but can be used on either static images or videos at inference time. As the volume of our training sets grows, more and more objects are seen moving, thus turning our method into unsupervised (or time-supervised) training to segment primary objects. The resulting system outperforms the state-of-the-art in both video object segmentation and salient object detection benchmarks, even when compared to methods that use explicit manual annotation. Code is available.  

1 Introduction

We propose a method to learn how to segment primary objects in images by processing videos without any human annotation. We define objects as portions of physical scenes that, at some point in time, are seen moving as a coherent whole in an image sequence. While such motion is needed to bootstrap the training process (bottom-up), learned objects reinforce the model to the point where previously unseen objects can be segmented in static images through a “top-down” model.

Objects that have never been seen moving are technically not captured in our definition. Instead, they are considered part of whatever background they are attached to. However, as the set of videos grows, more and more objects are seen moving, thus improving our method’s ability to detect and segment objects in static images (Tab. 3). Since our method requires relative motion between objects and camera, which happens over time, we refer to it as “time-supervised.” Our method would not work if only presented with a photo collection of individual images.

Our bottom-up (or “dynamic”) branch of the model is trained to minimize the mutual information between partitions of the motion field, inspired by [63], while explicitly enforcing temporal consistency and shape regularity. This model provides evidence, in the form of pseudo-labels, to train a “static” branch to segment objects in a top-down fashion (Fig. 1). In the absence of a reliable confidence measure, the use of pseudo-labels can be counter-productive. However, we exploit the approximation of the mutual information computed as part of the bottom-up module as a confidence measure to inform the weight of the pseudo-labels. The trained static model is fed back to the bottom-up module as a regularizer, reinforcing the latter. The same architecture is used during inference, except that in the case of stills, the dynamic part is bypassed. The outcome of inference is a binary mask indicating the presence of any object in the scene.

* These authors contributed equally to this work.
1https://github.com/blai88/unsupervised_segmentation
Figure 1: System overview. Dynamic: motion detection module as described in Sec. 3.1. Static: static object module as described in Sec. 3.2. The training iterates between these two modules. Especially, when $\chi$ (static object model) is trained, it can be used as a top-down objectness prior to further promote the performance of the motion segmentation network $\phi$ through the feedback loop. $\psi$ is the adversarial inpainting network that enforces minimal mutual information between motion field partitions (two $\psi$’s are identical), and losses are represented using dashed lines/boxes.

One could argue that every pixel in the image back-projects to something in space that we could call an object. Thus, the output of the static model would be a constant mask covering the entire image. In practice, the biases in the datasets used for training determine the scale and typical motion of objects that informs what is labeled as such. We call these “primary objects” following the nomenclature in [67, 22, 26]. For instance, in consumer videos, seldom the resolution and the motion is such that every leaf of a tree is distinctly visible and segmentable by the bottom-up stage. Thus, the entire tree, standing against the background of grass and sky, would end up being a primary object. So, the definition of objects is conditioned on the training data and, in particular, the scale and distribution of size and relative motion of primary objects.

In summary, our main contributions are 1) a temporal constraint and a learned shape prior that improves prior work on motion-based segmentation based on mutual information minimization; 2) an efficient adaptive training scheme to learn a static object detection that reduces artifacts due to lack of informative motions; 3) a method to exploit the top-down object model as a generic objectness prior to further improve the bottom-up perception module in a feedback loop. The resulting primary object segmentation method outperforms the state-of-the-art in both video object segmentation benchmarks and static salient object detection benchmarks, despite being fully unsupervised. In addition to outperforming existing unsupervised methods by 10.1% on average, it also edges out supervised methods by 4.9% on average, on the benchmarks, despite requiring no explicit supervision.

2 Related Work

Motion segmentation aims to identify independently moving objects in a video and separating them from the context. Background subtraction assumes that motion causes intensity change [53, 10, 46], while scene dynamic models are proposed to compensate for the camera motion [40, 44, 47]. One could directly segment or cluster pixel-wise motion vectors [59, 50, 27, 43], but this approach is prone to errors due to noisy motion and occlusion. To reduce noise, methods also employ pixel trajectories accumulated over multiple frames for segmentation [3, 24, 61, 48]. In addition to the difficulty in solving complicated optimizations, it is hard to get features that accurately distinguish moving objects due to the discriminability-robustness trade-off. On the other hand, top performing deep models largely rely on manual pixel-level annotations to learn features for motion segmentation [14, 55, 60, 17, 9]. To overcome the scarcity of annotated data and to eliminate the inherent trade-offs common to unsupervised models, [63] proposes contextual information separation in which a segmentation network minimizes the mutual information between the motion field inside and outside the mask. Due to the lack of temporal consistency and shape regularity, [63] is sensitive to perturbations in the motion (Fig. 4), moreover, [63] can not detect stationary objects as it fully relies on motion information (Fig. 5).

Saliency prediction aims to detect the most salient objects in the scene and to determine their spatial extent. The contrast of image regions to their context is an important cue for saliency prediction
that can be computed either locally [25, 20, 39] or globally [66, 1, 7], or at multiple scales [34, 36].

[16] constructs the saliency map from the spectral residual and [49] performs low-rank matrix decomposition to detect salient objects. Despite the advancement in the optimizations for salient region segmentation [8, 13, 29], the quality of the predicted saliency depends highly on hand-crafted features. Currently, the best performing methods employ deep neural networks trained on labeled datasets [69, 38, 35, 33, 32]. Later, [68, 41] propose training deep networks using the pseudo-labels generated by conventional unsupervised methods. Recently, [6, 2] propose unsupervised adversarial salient object discovery models on single images. Due to enormous variations in object appearance, these models are always hard to train and require extremely large training set.

Video object segmentation Due to the limited space, we only discuss the unsupervised methods which are more related to our work. However, note that the definition of “unsupervised” in video object segmentation only reflects whether annotations are used during testing, different from our common understanding that no annotation should be involved in any case. To segment moving objects, [55] trains a network to directly output the segmentation from motion, which is then augmented by an appearance channel in [21]. Further, [72] improves the segmentation by incorporating salient motion detection with object proposals. With region augmentation and reduction, [26] segments video objects based on the recurrence property of primary object. A pyramid dilated bidirectional ConvLSTM is proposed in [52] to extract spatial features at multiple scales, and [37] introduces a global co-attention mechanism to capture scene context. Interestingly, [70] proposes an architecture that allows interaction between motion and appearance during the encoding process, while [65] focuses on learning discriminative features for saliency propagation. Moreover, human attention is integrated in [58] to provide extra supervisions for video object segmentation. For clarity, we compare to these methods with an explicit list of the annotations used in the training phase.

Video:
Motion Fields:
Image I
~ P( u|I)

Figure 2: A single image I, renders the possible motion fields u of an independently moving object as if they are sampled from the conditional distribution p(u|I).

Figure 3: With image I in Fig. 2 only the mask in (c) minimizes mutual information for both the object and the background.

3 Method

An object, which is detachable from its surrounding, renders relative motion against its background. The only principle underlying our work is that objects move independently, thus the context should be maximally uninformative of the motion of objects. Here, we utilize the independence principle directly by minimizing the mutual information between the motions of an object and its context in Sec. 3.1. In contrast to the Contextual Information Separation (CIS) criteria [63], we enforce that the segments are both compatible with a learned shape prior and temporally consistent within a sequence. In Sec. 3.2, we instantiate another pathway that enables perception of stationary objects utilizing the detection of moving ones. Further, the interaction between motion detection and static perception is described in Sec. 3.3. The overall structure of our method is described in Fig. 1.

3.1 Mutual Information Minimization with Temporal Consistency and Shape Prior

Given an image I ∈ R^{H×W×C}, the motion field u defined on I is a random variable distributed according to p(u|I), which is determined by a dataset of image sequences D = {I^t}^t≤T, where N is the cardinality of the dataset and T is the maximum number of images in a sequence. Particularly, for an instance u ∈ R^{H×W×2}, sampled from p(u|I), there exists an image I such that for any pixel x on I, I(x) = I(x + u(x)) holds up to noise and occlusions, as shown in Fig. 2.

To detect objects that move independently in the scene, a motion segmentation network φ should generate masks m = φ(u) ∈ {0, 1}^{H×W}, such that the motion inside the mask m ⊙ u and outside
the mask \( (1 - m) \odot u \) are mutually independent conditioned on the image \( I \). More explicitly, the
conditional mutual information \( I(m \odot u, (1 - m) \odot u|I) \) should be minimal. Since the mutual
information measures the difference between the Shannon entropy of the inside and its conditional
entropy on the outside, simply minimizing the mutual information yields a trivial solution (empty set).
One solution is to normalize the conditional mutual information by the entropy \( H(m \odot u|I) \),
which is equivalent to maximizing the un-informativeness measured by the ratio on the right. Again,
this ratio can be maximized by simply taking the whole image domain, which results in over detection
(Fig. 3 (a)). Thus, if the detection is not accurate, either over or under detection (Fig. 3), the context
will be rendered informative, vice versa. Therefore it is necessary to use a symmetric term:

\[
L(m; I) = \frac{H(m \odot u|(1 - m) \odot u, I)}{H(m \odot u|I)} + \frac{H((1 - m) \odot u|m \odot u, I)}{H((1 - m) \odot u|I)}
\]

(2)

whose value is upper bounded by 2, and can be maximized only when \( m \) accurately segments the
object. By assuming gaussian conditionals \([63]\), Eq. (2) can be made computable with an adversarial
inpainting network \( \psi \) that computes the conditional means, e.g., \( \psi(m, (1 - m) \odot u, I) \) estimates the
conditional mean of \( m \odot u \) under \( p(m \odot u|(1 - m) \odot u, I) \). Both \( \phi \) and \( \psi \) can be trained adversarially:

\[
\max_{\phi} \min_{\psi} \mathcal{L}_A(\phi, \psi; I) = \frac{\sum_{u \sim P(u|I)} \| m \odot u - \psi(m, (1 - m) \odot u, I) \|}{\sum_{u \sim P(u|I)} \| m \odot u \| + \epsilon} + \frac{\sum_{u \sim P(u|I)} \| (1 - m) \odot u - \psi(1 - m, m \odot u, I) \|}{\sum_{u \sim P(u|I)} \| (1 - m) \odot u \| + \epsilon}
\]

(3)

with \( m = \phi(u) \), \( \odot \) the Hadamard product, and \( \| \cdot \| \) the \( l^2 \)-norm. The constant \( 0 < \epsilon \ll 1 \) is to
prevent numerical instability, and \( \psi(m, \emptyset, I) \) is default to zeros.

Since Eq. (3) characterizes moving objects solely using motion, it is sensitive to variations in the
motion field, resulting in label flipping shown in Fig. 4. Moreover, the lack of constraint on the shape
of detected objects could result in highly irregular segments due to failures of motion estimation
(Fig. 5). To resolve these issues, we introduce a temporal constraint to reduce the instability in the
motion segmentation. Further, we explicitly learn a prior on possible shapes of objects in the scene.

![Figure 4: Temporal consistency constraint prevents label flipping.](image)

**Temporal Consistency.** Given two consecutive images \( I^1, I^2 \) from the same video sequence, we
can compute the forward and backward optical flow \( u^{12}, u^{21} \), and the predicted masks \( m^1 = \phi(u^{12}), m^2 = \phi(u^{21}) \). We would like the individually predicted masks to be temporally consistent, in
the sense that if we deform one onto the other, the two should look similar as they are the projections of
the same object. Thus, we penalize the following warping difference to enforce temporal consistency:

\[
L_{TC}(\phi; u^{12}, u^{21}) = \sum_{x \not\in o} |m^1(x) - m^2(x + u^{12}(x))| + |m^2(x) - m^1(x + u^{21}(x))|
\]

(4)

with \( x \) the pixel location, and \( o \) the union of occlusions within the image domain, which can be easily
estimated using the forward-backward identity criterion \([19]\). The reasoning is that inconsistencies of
the predictions should only be penalized in the co-visible region. Note, Eq. (4) effectively reduces instabilities in the motion predictions (Fig. 5) compared to Eq. (3).

Object Shape Prior. Objects encountered in our daily life are compact and multiply-connected, thus their projections on 2D can not be arbitrary. On the other hand, there is plenty of object shape models collected using RGB-D or 3D scans [4], which provides rich information on the shape space of the 2D segments. Our goal is to learn a differentiable prior on the projected shapes of general objects, such that the shapes of the motion detection (Eq. (3)) can be regularized. For this purpose, we adapt Conditional Prior Networks proposed in [64], and use it to learn a shape prior with a network $Q$, by minimizing the information regularized training loss:

$$\min_Q KL(p(m)\|Q(m)) + \beta I(e_Q(m),m)$$

with $e_Q$ the encoding part of $Q$. We render 2D projections using 3D models from ShapeNet [4] to synthesize the distribution of the 2D object shapes $p(m)$, more details in Suppmat. When trained, $Q(m)$ approximates $p(m)$ up to a scalar as indicated by the Kullback–Leibler divergence term, and the information regularizer prevents overfitting, which would result in an uninformative prior.

Motion Detection Loss: By imposing temporal consistency and the learned shape regularity, our model for independently moving object detection is summarized as following:

$$\max_{\phi} \min_{\psi} L_D(\phi,\psi; I_1, u_{12}, u_{21}, Q) = L_A(\phi,\psi; I_1) + \lambda_{TC} L_{TC}(\phi; u_{12}, u_{21}) + \lambda_Q \log(Q(m^1))$$

with $m^1 = \phi(u_{12})$, and $\lambda_{TC}$=-1.0, $\lambda_Q$=1e-5 the relative weights. Note $\lambda_{TC} < 0$ as $\phi$ maximizes $L_D$.

Figure 5: Motion model v.s. static object model. After learning from the noisy motion detections ($\phi(u)$), the static object model $\chi(I)$ improves over $\phi$ on all cases where the motion is noisy (a), the object is partially static (b) or fully static (c).

3.2 Learning Prior on Stationary Objects

In Sec. 3.1 we describe a model that detects moving objects in a temporally coherent and geometrically regularized manner. However, what if the objects stop moving? This causes difficulty for the motion detection network $\phi$, which relies heavily on motion to signal the existence of an object (Fig. 5). Also note that in Fig. 5, the motion varies between frames, but the appearance of an object is temporally persistent, and when motion fades away, the image array still depicts the same object. Thus, we propose to train a static object model that segments objects only from image arrays to complement the detection when there is no prominent motion. More precisely, we train a segmentation network $\chi$ to detect general objects from a single image, utilizing the output of $\phi$ as the pseudo labels. We can train $\chi$ in supervised manner by maximizing the F-measure:

$$F_\alpha(\chi(I),\overline{m}) = (1+\alpha^2)\frac{\rho(\chi(I),\overline{m})\gamma(\chi(I),\overline{m})}{\alpha^2\rho(\chi(I),\overline{m}) + \gamma(\chi(I),\overline{m})}$$

with $\rho, \gamma$ the precision and recall between the predicted segmentation $\chi(I)$ and the motion mask $\overline{m}$ from $\phi$ ($\alpha^2$ is default to 1.5 if not explicitly mentioned). However, directly learning from all motion masks without discernment will be counter-productive, as the motion predictions could be quite noisy or even erroneous when the relative motion is not informative (Fig. 5). Thus we propose to use the loss $L_\chi$ (Eq. 5) as an indicator of the reliability of the motion predictions: if the motion is not discriminative, the reconstruction of the motion from the context will be easy, thus $L_\chi$ will be small,
vice versa, if the motion is distinctive, $L_A$ will be large due to the difficulty in a good reconstruction. More explicitly, we propose the following static model training loss for $\chi$:

$$L_A(\chi; I, \overline{m}, \psi, \chi') = \max (L_A(\overline{m}, \psi; I) - L_A(\chi'(I), \psi; I) - \delta, 0) F_u(\chi(I), \overline{m}) + \lambda_F F_u(\chi(I), \chi'(I))$$  \hspace{1cm} (8)$$

Note that, the pseudo label $\overline{m}$ is only effective when it has a larger $L_A$ than the one predicted by $\chi'$, which is a copy of an earlier $\chi$. In other words, if the reliability of the motion prediction measured by $L_A$ is not high enough, $\chi$ retains its own prediction. We set $L_A(\chi'(I), \psi; I)=0$ and $\lambda_F=0$ the first time $\chi$ is trained, then $\lambda_F=1.0$. When trained, $\chi$ learns a model of objects based on their appearance, so we would expect $\chi$ to detect stationary objects which have been seen moving before. Indeed, we find that $\chi$ is able to detect static object that has never been observed moving as shown in Fig. 5, which confirms that general concept of objects can be learned through the observations of moving ones.

### 3.3 Joint Perception of Objects from Motion and Appearance

Once the static object model has been learned by $\chi$, it can be used to modulate the detection in general scenes, even where the motion of the objects is unknown. It is possible that the predictions of $\phi$, which employs motion information are imperfect. The output of $\chi$ can then provide complementary information to strengthen the detection model. Similar to the shape prior in Eq. (6) that regularizes the shape of the segments, we feed $\chi$ back into the training of the bottom-up module as a more sophisticated prior of general objectness based on the photometric information. The joint detection model with the objectness prior is:

$$\max_{\phi} \min_{\psi} L_J(\phi, \psi; I^1, u^{12}, u^{21}, Q, \chi) = L_J(\phi, \psi; I^1, u^{12}, u^{21}, Q) + \lambda_{obj} F_u(\phi(u^{12}), \chi(I^1))$$  \hspace{1cm} (9)$$

with $L_J$, the adversarial motion detection loss in Eq. (6), and the second term measures the similarity between the motion prediction and the static objectness prior, with $\lambda_{obj}=1.0$. Besides learning from motion information to detect moving objects, $\phi$ is now able to leverage photometric cues that facilitate the detection under circumstances where objects become stationary or move extremely slowly. Moreover, improved motion detection could yield better pseudo-labels that help training a more accurate static object model, which can then be used to facilitate the learning of the former in a feedback loop. The overall training procedure is presented in Algorithm 1.

| Algorithm 1: The overall training procedures |
|------------------------------------------------|
| **Result:** $\phi$: motion detection model; $\chi$: static object model; $\psi$: conditional inpainting network |
| Initialize $\phi, \psi$ by optimizing $L_J$ (Eq. (6)), set $k=0$; |
| while $k \leq 2$ do |
| $k = k+1$; |
| Update the static model $\chi$ by maximizing $L_A$ (Eq. (8)); |
| Update $\phi, \psi$ using the joint detection loss $L_J$ (Eq. (9)); |
| $|$ |

### 4 Implementation

**Motion Detection** $\phi$ uses the Deeplab architecture [5]. We initialize the ResNet101 backbone with pretrained weights from Imagenet [68, 41, 65], while the other layers are randomly initialized. $\phi$ takes as input the estimated flow between two randomly chosen frames from the same video with the maximum interval equals to three. Optical flow is estimated using PWCNet [54]. The output of $\phi$ is a two channel segmentation mask after the softmax activation, and the inference speed is 19 fps.

**Adversarial Inpainting** $\psi$: The flow inpainting network $\psi$ utilizes a siamese network architecture with two separate encoders. The first encoder takes the mask and the masked flow as input, while the second takes in the image as input. Skip connections between the two encoders and the joint decoder are enabled. The output is a two channel scalar field representing the inpainted flow.

**Static Object Model** $\chi$ also uses the Deeplab architecture, similarly, the ResNet101 backbone is initialized using pretrained weights on ImageNet. $\chi$ takes a single RGB image as input. The output is a two channel segmentation mask after the softmax activation. Note for ablation studies, we use the ResNet50 backbone to speed up the training. $\chi$ can run at 22fps.
Training Details: First, in the dynamic stage, we train the motion detection model $\phi$. Concretely, we train $\phi$ and $\psi$ adversarially. We alternate between updating $\phi$ for three steps and updating $\psi$ for one step, for 30 epochs. We use an Adam optimizer with $lr=1e-4$, $beta1=0.9$, and $beta2=0.999$. In the second stage, we train the static model and freeze the weights in the dynamic model. The static model is trained up to 15 epochs, using the Adam optimizer with $lr=2e-5$, $beta1=0.9$, and $beta2=0.999$. To generate the pseudo-labels used to bootstrap the static object model, we take the sum of the dynamic model’s predictions from six different time steps, i.e., sample the flow fields for a single image by varying the second image in a neighbourhood of plus and minus three. We then perform normalization such that the pseudo-label is a scalar field with values between 0 and 1.

5 Experiments

Datasets: For video object segmentation, we evaluate our method on three commonly used video object segmentation datasets: DAVIS [45], FBMS [42], and SegTrackV2 [30]. DAVIS consists of high-resolution videos (30 for training and 20 for validation) depicting the primary object moving in the scene with pixel-wise annotations for each frame. Compared to DAVIS, the image resolution of FBMS and SegTrackV2 is much lower. FBMS contains videos of multiple moving objects, providing test cases for multiple object segmentation. FBMS has sparsely annotated 59 video sequences, with 30 sequences for validation. SegTrackV2 contains 14 densely annotated videos. These videos constitute the only source of training data for our unsupervised motion perception module $\phi$.

To evaluate $\chi$ on static object segmentation, we employ three major saliency prediction datasets: MSRA-B [23] (5000 images), ECCSD [51] (1000 images) and DUT [62] (5168 images). All three datasets are annotated with pixel-wise labels for each image. These saliency datasets contain objects from a much broader span of categories such as road signs, statues, flowers, etc., that are never seen moving in the training videos. We evaluate the static object prior $\chi$ learned from only video objects on the saliency benchmarks, to check its transferability to different instances from seen categories and unseen categories.

5.1 Effectiveness of Temporal Consistency and Shape Prior

To verify the effectiveness of the temporal consistency constraint and the shape prior proposed in Sec. 3.1 for bottom-up motion detection, we compare to the baseline CIS [63] that trains a segmentation network using only Eq. (3). We train both CIS [63] and our model described in Eq. (6) on the unlabeled videos from DAVIS, and then test on the validation set of DAVIS. We also report the scores by directly applying the model trained on DAVIS to FBMS and SegTrackV2 in Tab. 1 to check the generalization on different domains. In this study, we evaluate the prediction from a single time-step. The performance is measured by mean-Intersection-over-Union (mIoU), and the relative weights used in our model are $\lambda_{TC}=1.0$, $\lambda_{Q}=1e-5$. As shown in Tab. 1, our motion model (Eq. (6)) consistently outperforms CIS (Eq. (3)) on all three video object segmentation benchmarks.

| Table 2: Adaptive Bootstrapping | Table 3: $\chi$ Improves over Time |
|---------------------------------|-----------------------------------|
|                                 | DA VIS   | FBMS    | SegTV2  | # of videos | 12   | 36   | 60   | 84   | 108  |
| $\chi(F_{\alpha})$             | 73.8     | 65.5    | 65.7    | mIoU       | 48.3 | 52.4 | 54.9 | 58.8 | 61.4 |
| $\chi(L_{\chi})$               | 78.2     | 68.7    | 69.3    | Std. Dev.  | 1.99 | 2.37 | 2.30 | 1.17 | 1.56 |

5.2 Effectiveness of the Adaptive Bootstrapping

Here we check the effectiveness of the adaptive bootstrapping scheme proposed in Sec. 3.2. We first train a motion model $\phi$ on the training data from DAVIS. Then we train two static models $\chi$: $\chi(F_{\alpha})$ using Eq. (7), the other $\chi(L_{\chi})$ using Eq. (8). We set $\delta$ in Eq. (8) to 0.2 and $\lambda_{F}$ to 1.0, which are fixed for the future experiments. We compare the performance of the static models on the DAVIS validation set in terms of mIoU. Further, we perform the same evaluation on both FBMS and SegTrackV2, and report the scores in Tab. 2. As shown, with the adaptive bootstrapping loss Eq. (8), the static object model $\chi$ consistently improves over its counter-part on the three benchmarks.
5.3 Static Model Improves with the Number of Training Videos

One characteristic of our method is that the static model $\chi$ improves over time as more and more objects are seen moving through the bottom-up motion detection module $\phi$. To verify, we construct a collection of videos $S$ by combining the three aforementioned video object segmentation datasets (in total there are 123 video sequences). We randomly partition them into 10 subsets $\{s_k\}_{k=1}^{10}$, each contains around 12 video sequences. Correspondingly, we train 5 static object models $\{\chi_i\}_{i=0}^{4}$ by performing Algorithm 1. The training set for each $\chi_i$ is $\{s_k\}_{k=2i+1}^{2i+2}$, such that $\chi_i$ with a larger $i$ is exposed to more video sequences. Each $\chi$ is evaluated on the union of the three saliency datasets mentioned above (in total 11,000 testing images). In Tab. 3, we report the performance of $\chi_i$’s, measured in terms of mIoU (with standard deviation computed across five runs). As shown in Tab. 3, when the number of the observed videos increases, the performance of the static object model also improves, which is consistently observed across multiple runs.

5.4 Video Object Segmentation Benchmarks

We evaluate the proposed model (Fig. 1) on the task of video object segmentation. We compare to top performing unsupervised and supervised methods. We train our motion model on the unlabeled videos with temporal consistency and shape prior constraints with $\lambda_{TC} = 1.0$ and $\lambda_Q = 1e-5$, and then train the static model following Algorithm 1. Since video object segmentation focuses on moving objects, we weight the predictions from the static object model with the predictions from the motion model to emphasize the detection of moving objects. CRF postprocessing is performed to get our final results. The performance is measured by mIoU. As shown in Tab. 4, our method achieves the top performance on all three video object segmentation benchmarks among fully unsupervised methods. To compare with methods that utilize manual annotations (Supervised), we finetune our model on the DAVIS training set with around 2000 annotations. We have also listed the number of annotations used by other supervised methods in Tab. 4. Again, our model achieves the top performance using the least amount of manual annotations among all the supervised methods.

Table 4: Quantitative Comparison on Video Object Segmentation Benchmarks.

| Supervised Methods | Unsupervised Methods |
|--------------------|----------------------|
| # Annot. | DAVIS | FBMS(J/F) | SegTV2 | DAVIS | FBMS | SegTV2 |
| MATNet [70] | 14,000 | 82.4 | 76.1 | — | ARF [26] | 76.2 | 59.8 | 57.2 |
| AnDiff [65] | 2,000 | 81.7 | 76.1 | 81.2 | ELM [28] | 61.8 | 61.6 | — |
| COSNet [37] | 17,000 | 80.5 | 75.6 | 81.2 | FST [43] | 55.8 | 47.7 | 47.8 |
| EPONet [11] | 2,000+ | 80.6 | — | 70.9 | NLC [12] | 55.1 | 51.5 | 67.2 |
| PDB [52] | 17,000 | 77.2 | 74.0 | 81.5 | 60.9 | SAGE [57] | 42.6 | 61.2 | 57.6 |
| LVO [55] | 2,000+ | 75.9 | 65.1 | 77.8 | 57.3 | STP [18] | 77.6 | 60.8 | 70.1 |
| FSEG [21] | 10,500 | 70.7 | 68.4 | 61.4 | CIS [63] | 71.5 | 63.6 | 62.0 |
| Ours | 2,000 | 82.8 | 75.8 | 82.0 | — | Ours | 80.0 | 73.2 | 74.2 |

5.5 Unsupervised Salient Object Detection

In Tab. 5, we evaluate the learned object prior on salient object detection in images. Note that the top performing methods on unsupervised salient object detection all rely on handcrafted methods either as the main procedure, or as a subprocess. Among all top performing ones, we are, to the best of our knowledge, the only one that does not rely on any handcrafted features. We refine the static model $\chi$ by performing CRF on its predictions, and by one round of self-training with the CRF refined masks. By leveraging the object prior learned through videos, we can approach and surpass the state-of-the-art. Even when compared with top performing supervised methods (DSS, NDF, SR in Tab. 5), our method still achieves competitive performance with no explicit annotation.

Table 5: Quantitative Comparison on Salient Object Detection.

| Method | # Annot. | DAVIS | FBMS(J/F) | SegTV2 | ARF [26] | ELM [28] | FST [43] | NLC [12] | SAGE [57] | STP [18] | CIS [63] | Ours |
|--------|----------|-------|----------|--------|---------|----------|----------|----------|----------|----------|----------|-----|
| MATNet [70] | 14,000 | 82.4 | 76.1 | — | ARF [26] | 76.2 | 59.8 | 57.2 |
| AnDiff [65] | 2,000 | 81.7 | 76.1 | 81.2 | ELM [28] | 61.8 | 61.6 | — |
| COSNet [37] | 17,000 | 80.5 | 75.6 | 81.2 | FST [43] | 55.8 | 47.7 | 47.8 |
| EPONet [11] | 2,000+ | 80.6 | — | 70.9 | NLC [12] | 55.1 | 51.5 | 67.2 |
| PDB [52] | 17,000 | 77.2 | 74.0 | 81.5 | 60.9 | SAGE [57] | 42.6 | 61.2 | 57.6 |
| LVO [55] | 2,000+ | 75.9 | 65.1 | 77.8 | 57.3 | STP [18] | 77.6 | 60.8 | 70.1 |
| FSEG [21] | 10,500 | 70.7 | 68.4 | 61.4 | CIS [63] | 71.5 | 63.6 | 62.0 |
| Ours | 2,000 | 82.8 | 75.8 | 82.0 | — | Ours | 80.0 | 73.2 | 74.2 |

Figure 6: Failure case: multiple salient objects appear in the same image are all highlighted.
Table 5: Quantitative Comparison on Saliency Benchmarks.

|       | DSS | NDF | SR | RBD | DSR | HS | USD | DUSPS | Ours |
|-------|-----|-----|----|-----|-----|----|-----|-------|------|
| ECCSD | 87.9| 89.1| 82.6| 65.2| 63.9| 62.3| 87.8| 87.4  | 88.1 |
| MSRA-B| 89.4| 89.7| 85.1| 75.1| 72.3| 71.3| 87.7| 90.3  | 89.7 |
| DUT   | 72.9| 73.6| 67.2| 51.0| 55.8| 52.1| 71.6| 73.6  | 73.9 |

6 Discussion

We have presented a method to learn segmenting objects in images that exploits temporal consistency in their motion, observed in training videos, for bootstrapping a top-down model. The definition of what constitutes an object is implicit in the method and in the datasets used for training. This may appear to be a limitation, as training on different datasets may yield different outcomes. However, what constitutes an object, or even a segment of an image, is ultimately not objective: In Fig. 4, is the object a person? Or the motorcycle she is riding? The union of the two? The helmet she is wearing? All of the above? We let the evidence bootstrap the definition: If the motion at the resolution of the first video shows the human and bike moving as a whole, we do not know any better than consider them one object. If, in later video, a human is seen without a motorcycle, she will be learned as an object thereon after. We use the term “primary” to characterize this aspect of our model. Admittedly, our model does not capture the fact that a proper object model should segment instances and enable multiple membership for each point: A pixel on the helmet is part of the object, but also of the person, and the rider, and so on. We also do not exploit side information from other modalities, or human annotations, when available. Nonetheless, despite the complete absence of annotation requirements, our method edges out methods that exploit manual annotation, so we believe it to be a useful starting point for further development of more complete object segmentation methods.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In 2009 IEEE conference on computer vision and pattern recognition, pages 1597–1604. IEEE, 2009.
[2] Adam Bielski and Paolo Favaro. Emergence of object segmentation in perturbed generative models. In Advances in Neural Information Processing Systems, pages 7254–7264, 2019.
[3] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In European conference on computer vision, pages 282–295. Springer, 2010.
[4] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository, Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.
[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2017.
[6] Mickael Chen, Thierry Artieres, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In Advances in Neural Information Processing Systems, pages 12705–12716, 2019.
[7] 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.
[8] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on pattern analysis and machine intelligence, 24(5):603–619, 2002.
[9] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting anything that moves. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 0–0, 2019.
[10] Ahmed Elgammal, Ramani Duraiswami, David Harwood, and Larry S Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. Proceedings of the IEEE, 90(7):1151–1163, 2002.
[11] Muhammad Faisal, Ijaz Akhter, Mohsen Ali, and Richard Hartley. Exploiting geometric constraints on dense trajectories for motion saliency. arXiv preprint arXiv:1909.13258, 2019.
[12] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting.
[13] Pedro P Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. International journal of computer vision, 59(2):167–181, 2004.
Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for

Alex Levinshtein, Adrian Stere, Kiriakos N Kutulakos, David J Fleet, Sven J Dickinson, and Kaleem

Margret Keuper, Bjoern Andres, and Thomas Brox. Motion trajectory segmentation via minimum cost

Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene

Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance

Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation

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

Serdar Ince and Janusz Konrad. Occlusion-aware optical flow estimation. IEEE Transactions on Image

Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene

Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation

Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised

Ping Hu, Gang Wang, Xiangfei Kong, Jason Kuen, and Yap-Peng Tan. Motion-guided cascaded refinement

Won-Dong Jang, Chulwoo Lee, and Chang-Su Kim. Primary object segmentation in videos via alternate

Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by

Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Unsupervised video object segmentation using

Feng Liu and Michael Gleicher. Region enhanced scale-invariant saliency detection. In 2006 IEEE

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised

Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object
detection with short connections. In Proceedings of the IEEE Conference on Computer Vision and Pattern

M Pawan Kumar, Philip HS Torr, and Andrew Zisserman. Learning layered motion segmentations of video.

Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation

Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense

Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised

Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation

Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár,

Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense

Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense

Feng Liu and Michael Gleicher. Region enhanced scale-invariant saliency detection. In 2006 IEEE

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised

Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense

Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Unsupervised video object segmentation using

Feng Liu and Michael Gleicher. Region enhanced scale-invariant saliency detection. In 2006 IEEE

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár,

Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense

Feng Liu and Michael Gleicher. Region enhanced scale-invariant saliency detection. In 2006 IEEE

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense

Feng Liu and Michael Gleicher. Region enhanced scale-invariant saliency detection. In 2006 IEEE

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense

Feng Liu and Michael Gleicher. Region enhanced scale-invariant saliency detection. In 2006 IEEE

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense

Feng Liu and Michael Gleicher. Region enhanced scale-invariant saliency detection. In 2006 IEEE

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Feng Liu and Michael Gleicher. Region enhanced scale-invariant saliency detection. In 2006 IEEE

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving

Feng Liu and Michael Gleicher. Region enhanced scale-invariant saliency detection. In 2006 IEEE
[37] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3623–3632, 2019.

[38] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In Proceedings of the IEEE Conference on computer vision and pattern recognition, pages 6609–6617, 2017.

[39] Yu-Fei Ma and Hong-Jiang Zhang. Contrast-based image attention analysis by using fuzzy growing. In Proceedings of the eleventh ACM international conference on Multimedia, pages 374–381, 2003.

[40] Anurag Mittal and Nikos Paragios. Motion-based background subtraction using adaptive kernel density estimation. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., volume 2, pages II–II. Ieee, 2004.

[41] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummid, Nghung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepupps: Deep robust unsupervised saliency prediction via self-supervision. In Advances in Neural Information Processing Systems, pages 204–214, 2019.

[42] 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.

[43] 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.

[44] Kedar Patwardhan, Guillermo Sapiro, and Vassilios Morellas. Robust foreground detection in video using pixel layers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(4):746–751, 2008.

[45] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 724–732, 2016.

[46] Richard J Radke, Srinivas Andra, Omar Al-Kofahi, and Badrinath Roysam. Image change detection algorithms: a systematic survey. IEEE transactions on image processing, 14(3):294–307, 2005.

[47] Yaser Sheikh, Omar Javed, and Takeo Kanade. Background subtraction for freely moving cameras. In 2009 IEEE 12th International Conference on Computer Vision, pages 1219–1225. IEEE.

[48] Jianbing Shen, Jiaying Fu, and Ling Shao. Submodular trajectories for better motion segmentation in videos. IEEE Transactions on Image Processing, 27(6):2688–2700, 2018.

[49] Xiaohui Shen and Ying Wu. A unified approach to salient object detection via low rank matrix recovery. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 853–860. IEEE, 2012.

[50] 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.

[51] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. IEEE transactions on pattern analysis and machine intelligence, 38(4):717–729, 2015.

[52] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In Proceedings of the European Conference on Computer Vision (ECCV), pages 715–731, 2018.

[53] Chris Stauffer and W. Eric L. Grimson. Learning patterns of activity using real-time tracking. IEEE Transactions on pattern analysis and machine intelligence, 22(8):747–757, 2000.

[54] 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.

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

[56] Tianjian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In Proceedings of the IEEE International Conference on Computer Vision, pages 4019–4028, 2017.

[57] 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.

[58] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3064–3074, 2019.

[59] Yair Weiss. Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 520–526. IEEE, 1997.

[60] Huaxin Xiao, Jiafeng Fu, Guosheng Lin, Yu Li, and Maojun Zhang. Monet: Deep motion exploitation for video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1140–1148, 2018.

[61] Hongteng Xu, Yang Zhou, Weiyao Lin, and Hongyuan Zha. Unsupervised trajectory clustering via adaptive multi-kernel-based shrinkage. In Proceedings of the IEEE International Conference on Computer Vision, pages 4328–4336, 2015.
[62] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3166–3173, 2013.

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

[64] Yanchao Yang and Stefano Soatto. Conditional prior networks for optical flow. In Proceedings of the European Conference on Computer Vision (ECCV), pages 271–287, 2018.

[65] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In Proceedings of the IEEE International Conference on Computer Vision, pages 931–940, 2019.

[66] Yun Zhai and Mubarak Shah. Visual attention detection in video sequences using spatiotemporal cues. In Proceedings of the 14th ACM international conference on Multimedia, pages 815–824, 2006.

[67] Dong Zhang, Omar Javed, and Mubarak Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 628–635, 2013.

[68] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtaish 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.

[69] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 202–211, 2017.

[70] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation.

[71] 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.

[72] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. Unsupervised online video object segmentation with motion property understanding. IEEE Transactions on Image Processing, 29:237–249, 2019.

[73] 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.