# DyStaB: Unsupervised Object Segmentation via Dynamic-Static Bootstrapping*

Yanchao Yang[†]
Stanford University
yanchaoy@cs.stanford.edu

Brian Lai[†]
UCLA Vision Lab
b4lai@g.ucla.edu

Stefano Soatto
UCLA Vision Lab
soatto@ucla.edu

## Abstract

*We describe an unsupervised method to detect and segment portions of images of live scenes that, at some point in time, are seen moving as a coherent whole, which we refer to as objects. Our method first partitions the motion field by minimizing the mutual information between segments. Then, it uses the segments to learn object models that can be used for detection in a static image. Static and dynamic models are represented by deep neural networks trained jointly in a bootstrapping strategy, which enables extrapolation to previously unseen objects. While the training process requires motion, the resulting object segmentation network can be used on either static images or videos at inference time. As the volume of seen videos grows, more and more objects are seen moving, priming their detection, which then serves as a regularizer for new objects, turning our method into unsupervised continual learning to segment objects. Our models are compared to the state of the art in both video object segmentation and salient object detection. In the six benchmark datasets tested, our models compare favorably even to those using pixel-level supervision, despite requiring no manual annotation.*

## 1. Introduction

The ability to segment the visual field into coherently moving regions is among the traits most broadly shared among visual animals [2, 17, 18]. Even camouflaged, moving objects are easy to spot (Fig. 1). During early development, humans spend considerable amounts of time interacting with a single moving object before losing interest [4], which may help prime object models and learn invariances [59]. In contrast, the mature visual system can learn an object with a few static examples; objects do not need to move in order to be detected. This suggests using motion as a cue to bootstrap object models that can be used for detection in static images, with no need for explicit supervision. Ob-
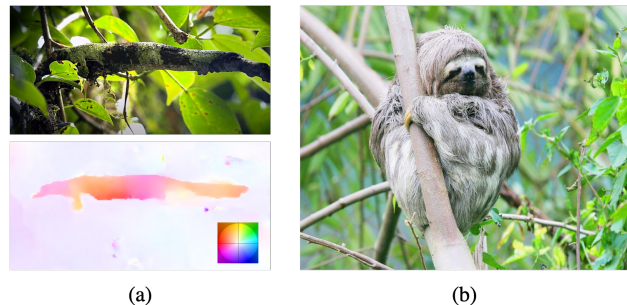


Figure 1. Dynamic-static bootstrapping. (a) A lizard is hard to detect when still thanks to camouflage (top left). However, it is easy to see once it moves (bottom left; optical flow visualized using the inset color wheel). Once learned the lizard, a never-before-seen sloth (b) can be easily detected in a static image, exploiting the static model learned from the moving lizard.

jects that have never been seen moving are considered part of whatever background they are part of, at least until they move. As time goes by, more and more objects are seen moving, thus improving one's ability to detect and segment objects in static images (Tab. 6). The more objects are bootstrapped in a bottom-up fashion, the easier they are to detect top-down, priming better motion discrimination, which in turn results in more accurate object detection. This synergistic loop gradually improves both the diversity of objects that can be detected and the accuracy of the detection.

We present a method to learn object segmentation using unlabeled videos, that at test time can be used for both moving objects in videos and static objects in single images. The method uses a motion segmentation module that performs temporally consistent region separation. The resulting motion segmentation primes a detector that operates on static images, and feeds back to the motion segmentation module, reinforcing it. We call this method Dynamic-Static Bootstrapping, or *DyStaB*. During training, the dynamic model minimizes the mutual information between partitions of the motion field, while enforcing temporal consistency, which yields a state-of-the-art unsupervised motion segmentation method. Putative regions, along with their uncertainty approximated during the computation of mutual information in the dynamic model, are used to train a static model. The
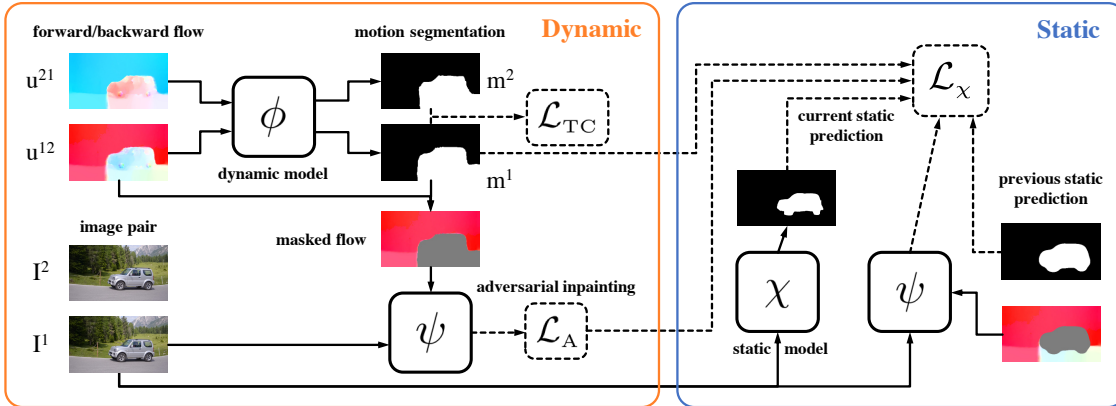
---

Figure 2. System overview. Dynamic: motion segmentation model described in Sec. 3.1; Static: object model described in Sec. 3.2. The training ("dynamic-static bootstrapping" in Sec. 3.3) iterates between these two models. In particular, once $\chi$ (static object model) is trained, it can be used as a "top-down" object prior to bias the motion segmentation network $\phi$ (dynamic model) in a 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.

static model is then fed back as a regularizer in a top-down fashion, completing the synergistic loop.

One might argue that every pixel in the image back-projects to something in space that we could call an object. However, the training data biases the process towards objects that exist at a scale that is detectable relative to the size of the pixel and the magnitude of the motion. For instance, individual leaves in an outdoor video might not be seen at a resolution that allows detecting them as independent objects. Instead, the tree may be detected as moving coherently. So, the definition of objects is conditioned on the training data and, in particular, the scale and distribution of their size and relative motion.

With this caveat, our contribution is two-fold: First, a deep neural network trained with unlabeled videos that achieves state-of-the-art performance in motion segmentation. It exploits mutual information separation and temporal consistency to identify candidate objects. Second, a deep neural network to perform object segmentation in single images, bootstrapped from the first. The static model uses as input both the output of the dynamic model and its uncertainty, to avoid self-learning. The two models are trained jointly in a synergistic loop. The resulting object segmentation models outperform the state of the art by 10% in average precision in both video and static object segmentation across six standard benchmarks. Despite not requiring any manual annotation, our method also outperforms recent supervised ones by almost 5% on average.

## 2. Related Work

**Motion segmentation** aims to identify independently moving regions in a video. Background subtraction assumes a static camera [60, 12, 52], while scene dynamic models compensate for camera motion [45, 49, 53]. One could directly segment or cluster pixel-wise motion vectors [69, 56, 31, 48], but this approach is prone to errors due to occlusions, singularities and discontinuities of the motion field. To increase robustness, some employ pixel trajectories accumulated over multiple frames [6, 27, 71, 54] or patches, losing discriminative power. The hard trade-off between discriminability and robustness is a key challenge in unsupervised motion segmentation. Manual pixel-level annotations help set the trade-off, but in a non-scalable manner [16, 63, 70, 21, 11]. Contextual Information Separation [76] aims to bypass this trade-off without the need for human annotation, using a segmentation network to minimize the mutual information between the inside and the outside of putative motion regions. In the absence of temporal consistency, this procedure can be sensitive to motion errors (Fig. 5). Furthermore, [76] cannot detect stationary objects as it relies on motion segmentation (Fig. 6). Note, [75] applies [76] with perceptual cycle-consistency to separating multiple objects, but they focus on learning object-centric representations.

**Saliency prediction** aims to detect the most salient objects relative to their background or "context." Saliency can be computed locally in a bottom-up fashion [29, 24, 44] or globally [78, 1, 9], at multiple scales [39, 41] and top-down [64]. [20] constructs the saliency map from the spectral residual and [55] performs low-rank matrix decomposition to detect salient objects. Separation can also be achieved by increasing the divergence of the feature distributions [28]. Despite advancements in the optimization for salient region segmentation [10, 15, 34], the quality of the predicted saliency depends highly on the features selected. Currently, the best performing methods employ deep neural networks trained on labeled datasets [80, 43, 40, 38, 37]. Recently, [8, 5] proposed unsupervised adversarial salient object dis-
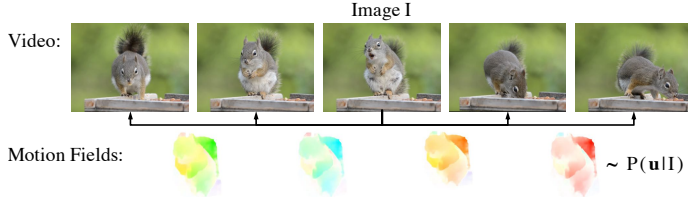
Figure 3. A single image I, renders possible motion fields $\mathbf{u}$ of an independently moving object as if they are sampled from the conditional distribution $p(\mathbf{u}|I)$. Thus, given the image, one can complete partial observations of the flow fields in Fig. 4.
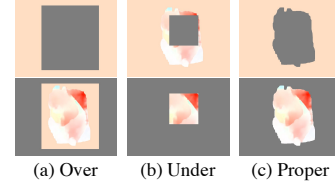


(a) Over     (b) Under     (c) Proper

Figure 4. With image I in Fig. 3, only the mask in (c) minimizes mutual information between flow fields of the object (inside mask) and the background (outside mask).

covery models in single images. Due to the variability of object appearance, these models are hard to train and the process is not scalable. On the other hand, [79, 46] propose training deep networks using the pseudo-labels generated by conventional unsupervised methods. While an improvement over adversarial methods, performance still hinges on human prior knowledge through the selection of handcrafted features. We wish to avoid specifying features by instead articulating a criterion, namely that anything similar to what we have previously detected as objects should be salient. The model trained with this simple criterion outperforms the state of the art on unsupervised saliency prediction.

**Video object segmentation** comprises a vast literature [13, 22, 33, 67, 68]; we focus on unsupervised methods that are more related to our work. It should be noted that the term "unsupervised" in video object segmentation only reflects absence of annotations at *test time*, whereas we reserve the name for methods that involve no manual annotation during training, as well as testing. To segment moving objects, [63] trains a network to directly output the segmentation from motion, which is then augmented by an appearance channel in [25]; [62] proposes a layered model for detachable objects using occlusion as primary cue. [83] incorporates salient motion detection with object proposals; [30] segments video into "primary" objects. A pyramid dilated bidirectional ConvLSTM is proposed in [58] to extract spatial features at multiple scales, and [42] introduces a global co-attention mechanism to capture scene context. [81] proposes an architecture that allows interaction between motion and appearance during the encoding process, while [77, 32] focuses on learning discriminative features for segmentation propagation and assumes the first frame been annotated.

## 3. Method

An object that is detached from its surroundings induces an independent motion when projected onto a moving image [3]. Here, we utilize this independence principle 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) criterion [76], we enforce that the separation is temporally consistent in a sequence, which

directly improves [76] by an average of 7%. In Sec. 3.2, we instantiate a static model that enables perception of stationary objects utilizing the detection of moving ones and the confidence measure computed from the mutual information. The interaction or mutual bootstrapping between motion segmentation and static perception is described in Sec. 3.3. The overall method is illustrated in Fig. 2.

### 3.1. Dynamic Model with Temporally Consistent Mutual Information Minimization

Given an image $I \in \mathbb{R}^{H \times W \times 3}$, the motion field $\mathbf{u}$ defined on I is a random variable distributed according to $p(\mathbf{u}|I)$, which is determined by a dataset of image sequences $\mathcal{D} = \{I_i^t\}_{i \leq N, t \leq T}$, where $N$ is the cardinality of the dataset and $T$ is the maximum number of images in a sequence. Particularly, for an instance $\mathbf{u} \in \mathbb{R}^{H \times W \times 2}$, sampled from $p(\mathbf{u}|I)$, there exists an image $\hat{I}$ such that for any pixel x on I, $I(x) = \hat{I}(x + \mathbf{u}(x))$ holds up to noise and occlusions, as shown in Fig. 3.

To detect objects that move independently in the scene, a motion segmentation network $\phi$ should generate masks $\mathbf{m} = \phi(\mathbf{u}) \in \{0, 1\}^{H \times W}$, such that the motions inside the mask $\mathbf{m} \odot \mathbf{u}$ and outside the mask $(1 - \mathbf{m}) \odot \mathbf{u}$ are mutually independent conditioned on the image I. More explicitly, the conditional mutual information $\mathbb{I}(\mathbf{m} \odot \mathbf{u}, (1 - \mathbf{m}) \odot \mathbf{u}|I)$ should be minimal. Since the mutual information measures the difference between the Shannon entropy of the inside and its conditional entropy (conditioned) 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 $\mathbb{H}(\mathbf{m} \odot \mathbf{u}|I)$,

$$\arg \min_{\mathbf{m}} \frac{\mathbb{I}(\mathbf{m} \odot \mathbf{u}, (1 - \mathbf{m}) \odot \mathbf{u}|I)}{\mathbb{H}(\mathbf{m} \odot \mathbf{u}|I)}$$
$$= \arg \max_{\mathbf{m}} \frac{\mathbb{H}(\mathbf{m} \odot \mathbf{u}|(1 - \mathbf{m}) \odot \mathbf{u}, I)}{\mathbb{H}(\mathbf{m} \odot \mathbf{u}|I)} \quad (1)$$

which is equivalent to maximizing the un-informativeness measured by the ratio on the right. Again, this ratio can be maximized by simply setting $\mathbf{m}$ as the whole image domain, which results in over detection (Fig. 4 (a)). Thus, if the detection is not accurate as in Fig. 4 (c), the context will be

$\mathbf{u}^{12}$
$\mathbf{u}^{21}$
$\mathbf{u}^{12}$
$\mathbf{u}^{21}$

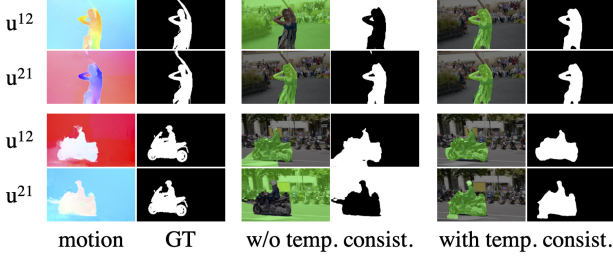motion    GT    w/o temp. consist.    with temp. consist.

Figure 5. Temporally consistent mutual information minimization prevents label flipping (2nd and 3rd columns), and also improves segmentation accuracy (the motorbike).

rendered informative, vice versa. Therefore it is necessary to add a symmetric term:

$$\mathcal{L}(\mathrm{m};\mathrm{I}) = \frac{\mathbb{H}(\mathrm{m} \odot \mathbf{u}|(1-\mathrm{m}) \odot \mathbf{u}, \mathrm{I})}{\mathbb{H}(\mathrm{m} \odot \mathbf{u}|\mathrm{I})}$$
$$+ \frac{\mathbb{H}((1-\mathrm{m}) \odot \mathbf{u}|\mathrm{m} \odot \mathbf{u}, \mathrm{I})}{\mathbb{H}((1-\mathrm{m}) \odot \mathbf{u}|\mathrm{I})} \quad (2)$$

whose value is upper bounded by 2, and can be maximized only when m accurately segments the object.

Similar to [76], we make the loss in Eq. (2) computable by assuming Gaussian conditionals, and instantiating an adversarial inpainting network $\psi$ that computes the conditional means, e.g., $\psi(\mathrm{m}, (1-\mathrm{m}) \odot \mathrm{u}, \mathrm{I})$ estimates the conditional mean of $\mathrm{m} \odot \mathbf{u}$ under $p(\mathrm{m} \odot \mathbf{u}|(1-\mathrm{m}) \odot \mathrm{u}, \mathrm{I})$. Both $\phi$ and $\psi$ can be trained adversarially:

$$\max_{\phi} \min_{\psi} \mathcal{L}_{\mathrm{A}}(\phi, \psi; \mathrm{I}) =$$
$$\frac{\sum_{\mathrm{u} \sim p(\mathbf{u}|\mathrm{I})} \|\mathrm{m} \odot \mathrm{u} - \psi(\mathrm{m}, (1-\mathrm{m}) \odot \mathrm{u}, \mathrm{I})\|}{\sum_{\mathrm{u} \sim p(\mathbf{u}|\mathrm{I})} \|\mathrm{m} \odot \mathrm{u}\| + \epsilon}$$
$$+ \frac{\sum_{\mathrm{u} \sim p(\mathbf{u}|\mathrm{I})} \|(1-\mathrm{m}) \odot \mathrm{u} - \psi(1-\mathrm{m}, \mathrm{m} \odot \mathrm{u}, \mathrm{I})\|}{\sum_{\mathrm{u} \sim p(\mathbf{u}|\mathrm{I})} \|(1-\mathrm{m}) \odot \mathrm{u}\| + \epsilon} \quad (3)$$

with $\mathrm{m} = \phi(\mathrm{u})$, and $\|\cdot\|$ the $l^2$-norm. The constant $0 < \epsilon \ll 1$ is to prevent numerical instability, and $\psi(\mathrm{m}, \emptyset, \mathrm{I})$ 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 and irregular segments due to failures of motion estimation shown in Fig. 5. To resolve these issues, we introduce a temporal consistency constraint to reduce the instability in the motion segmentation model.

**Temporal consistency.** Given two consecutive images $\mathrm{I}^1, \mathrm{I}^2$ from the same video sequence, we can compute the forward and backward optical flow $\mathrm{u}^{12}, \mathrm{u}^{21}$, and the predicted masks $\mathrm{m}^1 = \phi(\mathrm{u}^{12}), \mathrm{m}^2 = \phi(\mathrm{u}^{21})$. We would like the individually predicted masks to be temporally consistent, in the sense that if we deform one onto the other using the flow fields, 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:

$$\mathcal{L}_{\mathrm{TC}}(\phi; \mathrm{u}^{12}, \mathrm{u}^{21}) = \sum_{\mathrm{x} \notin \mathbf{o}} |\mathrm{m}^1(\mathrm{x}) - \mathrm{m}^2(\mathrm{x} + \mathrm{u}^{12}(\mathrm{x}))|$$
$$+ |\mathrm{m}^2(\mathrm{x}) - \mathrm{m}^1(\mathrm{x} + \mathrm{u}^{21}(\mathrm{x}))| \quad (4)$$

with x the pixel index, and $\mathbf{o}$ the union of occlusions within the image domain, which can be easily estimated using the forward-backward identity criterion [23]. The reasoning is that inconsistencies of the predictions should only be penalized in the co-visible region. Without introducing extra networks, Eq. (4) effectively reduces instabilities in the motion segmentation compared to Eq. (3), as shown in Fig. 5.

**Initial training for the dynamic model:** By enforcing temporal consistency, the dynamic model for independently moving object segmentation can be obtained through the following adversarial training:

$$\max_{\phi} \min_{\psi} \mathcal{L}_{\mathrm{D}}(\phi, \psi; \mathrm{I}^1, \mathrm{u}^{12}, \mathrm{u}^{21})$$
$$= \mathcal{L}_{\mathrm{A}}(\phi, \psi; \mathrm{I}^1) + \lambda_{\mathrm{TC}} \mathcal{L}_{\mathrm{TC}}(\phi; \mathrm{u}^{12}, \mathrm{u}^{21}) \quad (5)$$

with $\lambda_{\mathrm{TC}} = -0.1$ ($\lambda_{\mathrm{TC}} < 0$ as $\phi$ maximizes $\mathcal{L}_{\mathrm{D}}$). The effectiveness of optimizing Eq. (5) is also numerically demonstrated in Tab. 1. Next, we describe the confidence-aware training for the static object model.

### 3.2. Static Model with Confidence-Aware Update

In Sec. 3.1, we describe a model that detects moving objects in a temporally coherent manner. However, what if the objects stop moving or they have been moving in an indistinctive way? These present challenges for the dynamic model $\phi$, which relies on motion to signal the existence of an object (Fig. 6). Note in Fig. 6, 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.

Our position is that, once a moving object is detected, we ought to be able to find it even in a still image. Thus, we propose to train a static object model to complement the dynamic model when there is no significant motion. We could directly train a segmentation network $\chi$ to predict objects from a single image, utilizing the output of $\phi$ as the pseudo labels, by maximizing the F-measure [46] commonly used for salient object detection:

$$F_\alpha(\chi(\mathrm{I}), \overline{\mathrm{m}}) = (1 + \alpha^2) \frac{\rho(\chi(\mathrm{I}), \overline{\mathrm{m}})\gamma(\chi(\mathrm{I}), \overline{\mathrm{m}})}{\alpha^2 \rho(\chi(\mathrm{I}), \overline{\mathrm{m}}) + \gamma(\chi(\mathrm{I}), \overline{\mathrm{m}})} \quad (6)$$

with $\rho, \gamma$ the precision and recall between the prediction $\chi(\mathrm{I})$ and the motion mask $\overline{\mathrm{m}}$ generated by $\phi$. And $\alpha^2$ is default to 1.5 if not explicitly mentioned.

However, directly learning from all motion masks is counter-productive, as these are quite noisy especially when
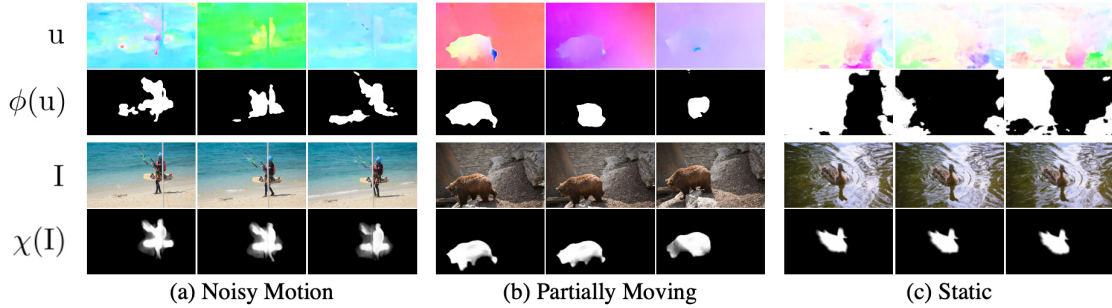
Figure 6. Comparison between the dynamic and static object models. After the first round of learning from noisy motion segmentations ($\phi(u)$, second row) with the proposed confidence-aware adaptation, the static object model ($\chi(I)$, fourth row) improves over the dynamic model on all cases where the motion is noisy (a), the object is partially static (b) or fully static (c).

motion is uninformative (Fig. 6). To address this issue, we propose to use the loss $\mathcal{L}_A$ (Eq. (3)) as a confidence measure on the reliability of the motion masks: *if the motion is uninformative, the reconstruction from the context will be accurate, thus $\mathcal{L}_A$ will be small and vice versa, if the motion is distinctive, $\mathcal{L}_A$ will be large due to a bad reconstruction* (See Fig. 7 and Fig. 8 for a demonstration.) We propose using the difference between the values of $\mathcal{L}_A$ to perform a confidence-aware adaptation via the following:

$$\mathcal{L}_\chi(\chi; I, \overline{m}, \chi') = \lambda_F F_\alpha(\chi(I), \chi'(I))$$
$$+ \max(\mathcal{L}_A(\overline{m}) - \mathcal{L}_A(\chi'(I)) - \delta, 0) F_\alpha(\chi(I), \overline{m}) \quad (7)$$

Note that the pseudo-label $\overline{m}$ is only effective when it has a larger $\mathcal{L}_A$ than the one predicted by $\chi'$, which is a copy of an earlier $\chi$. In other words, if the output of the dynamic model is not confident enough, $\chi$ retains its own prediction, and the first term (moving average) is to ensure that $\chi$ is updated smoothly. We set $\mathcal{L}_A(\chi'(I)) = 0$ and $\lambda_F = 0$ the first time $\chi$ is trained, then $\lambda_F = 1.0$.

When updated, $\chi$ 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. 6 (3rd column), which confirms that a general concept of objects can be learned through the observations of moving ones. Next we detail the proposed dynamic-static bootstrapping scheme for a continuous learning of objects.

### 3.3. Dynamic-Static Bootstrapping

Once the static object model is 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 also possible that the predictions of $\phi$, which employs motion information are imperfect. The output of $\chi$ can then provide complementary information to strengthen the dynamic model. We feed $\chi$ back into the training of the dynamic model as an experiential prior of objectness based on the photometric information.

The dynamic model reinforced by the objectness prior is:

$$\max_\phi \min_\psi \mathcal{L}_J(\phi, \psi; I^1, u^{12}, u^{21}, \chi) =$$
$$\mathcal{L}_D(\phi, \psi; I^1, u^{12}, u^{21}) + \lambda_{obj} F_\alpha(\phi(u^{12}), \chi(I^1)) \quad (8)$$

with $\mathcal{L}_D$ the adversarial motion segmentation loss in Eq. (5), and the second term measures the similarity between the motion mask and the static object prior ($\lambda_{obj} = 1.0$). 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 dynamic model could yield better pseudo-labels that help training a more accurate static object model (Eq. (7)), which can then be used to facilitate the learning of the former in a synergistic loop (Eq. (8)), thus the name "Dynamic-Static Bootstrapping." The overall training procedure is presented in Algorithm 1.

---

**Algorithm 1:** Dynamic-Static Bootstrapping

**Result:** $\phi$: dynamic model; $\chi$: static object model; $\psi$: conditional inpainting network

Initialize $\phi, \psi$ by optimizing $\mathcal{L}_D$ (Eq. (5)), set k=0;
**while** *k<3* **do**
    k = k+1;
    Update the static model $\chi$ using $\mathcal{L}_\chi$ (Eq. (7));
    Update the dynamic model $\phi$ using $\mathcal{L}_J$ (Eq. (8));

---

### 4. Implementation

**Dynamic model** $\phi$ uses the Deeplab architecture [7], with the initialization weights as in [79, 46]. $\phi$ 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 produced by PWCNet [61], which is trained on synthetic data. The output of $\phi$ is a two channel softmax score. In total, $\phi$ has 23M trainable parameters and can run in 19 fps during inference.
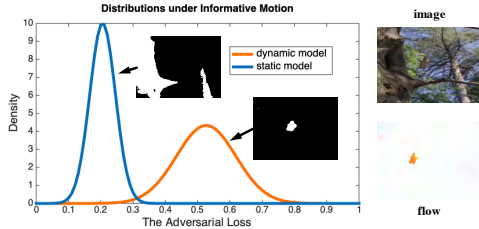
Figure 7. Distribution of $\mathcal{L}_A$ computed using the output of the dynamic model $\phi$ (orange) and the static model $\chi$ (blue) on the birdfall sequence, which shows that $\mathcal{L}_A$ is a good indicator of reliable masks when the motion field is informative, as verified by the gap between the distributions of $\mathcal{L}_A$.



Figure 8. Distribution of $\mathcal{L}_A$ for the dynamic model $\phi$ (orange) and the static model $\chi$ (blue) on the frog sequence, which shows that when the motion is not informative or erroneous, the static model can work better, thus the difference in the values of $\mathcal{L}_A$ is small.

**Static model** $\chi$ uses similar architecture as the dynamic model, but takes a single RGB image as input. The output is also a two channel softmax score. The total number of parameters is 21 M, and $\chi$ can run at 22 fps at inference.

**Training details.** For the initial training of the dynamic model $\phi$, we alternate between updating $\phi$ for three steps and updating $\psi$ for one step, up to 30 epochs on the training set of each dataset, using an Adam optimizer with lr=1e-4, beta1=0.9, and beta2=0.999. The static model is then trained up to 15 epochs, using an Adam optimizer with lr=2e-5, beta1=0.9, and beta2=0.999. As described in Algorithm 1, the dynamic-static bootstrapping runs for three iterations. On average, each iteration takes six hours to converge, and the whole training procedure can be finished within a day.

## 5. Experiments

**Datasets:** For *video object segmentation*, we train and test our model on three commonly used video object segmentation datasets: DAVIS [50], FBMS [47], and SegTrackV2 [35] (Tab. 3). We also test our trained model on two other datasets, DAVIS17 [51] and Youtube-VOS [72], to check how well our method generalizes (Tab. 7).

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. 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 $\phi$. Youtube-VOS contains 4,453 videos and 94 object categories, and DAVIS17 consists of 150 videos.

To evaluate $\chi$ on static object segmentation, we test on three major saliency prediction datasets: MSRA-B [26] (5000 images), ECCSD [57] (1000 images) and DUT [74] (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
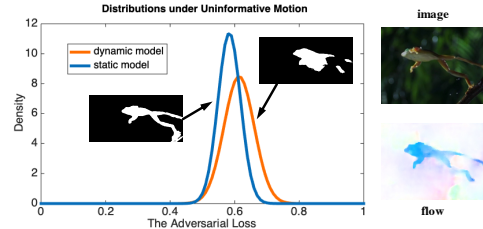
in the training videos We evaluate the static object model learned from only video objects on these saliency benchmarks, to check its transferability to different instances from seen categories and unseen categories.

### 5.1. Effectiveness of Temporally Consistent Mutual Information Minimization

To verify the effectiveness of the temporally consistent mutual information minimization proposed in Sec. 3.1 for bottom-up motion segmentation, we compare to the baseline CIS [76] that trains a segmentation network using only Eq. (3). We train both CIS [76] and our model described in Eq. (5) 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. The performance is measured by mean-Intersection-over-Union (mIoU), and the relative weights used in our model are $\lambda_{\text{TC}} = 1.0$. As shown in Tab. 1, our dynamic model (Eq. (5)) consistently outperforms CIS (Eq. (3)) on all three video object segmentation benchmarks by 7%, which confirms that temporal consistency is a critical component in our dynamic model.

|          | DAVIS | FBMS | SegTV2 |
|----------|-------|------|--------|
| CIS [76] | 59.2  | 36.8 | 45.6   |
| Ours     | **62.4** | **40.0** | **49.1** |

Table 1. Our temporally consistent dynamic model v.s. CIS [76].

### 5.2. Effectiveness of Confidence-Aware Adaptation

Here we check the effectiveness of the confidence-aware adaptation scheme proposed in Sec. 3.2 in precluding counter-productive self-learning. 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. (6) (not confidence-aware), the other $\chi(\mathcal{L}_\chi)$ using Eq. (7) (confidence-aware). We set $\delta$ in Eq. (7) to 0.2 and $\lambda_F$ to 1.0, which are fixed for the future experiments. We compare the performance of
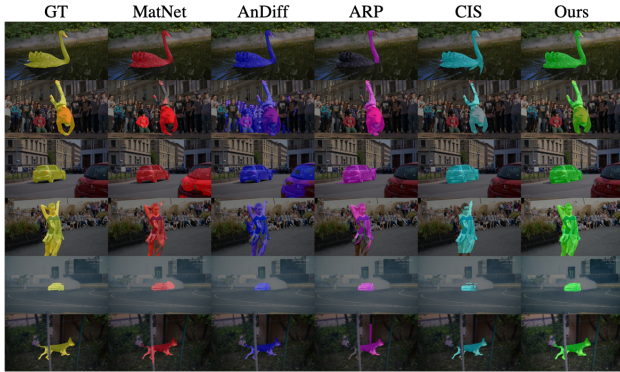
Figure 9. Qualitative comparison to the top two methods from each category on the DAVIS benchmark. MatNet [81], AnDiff [77] (supervised) and ARP [30], CIS [76] (unsupervised).
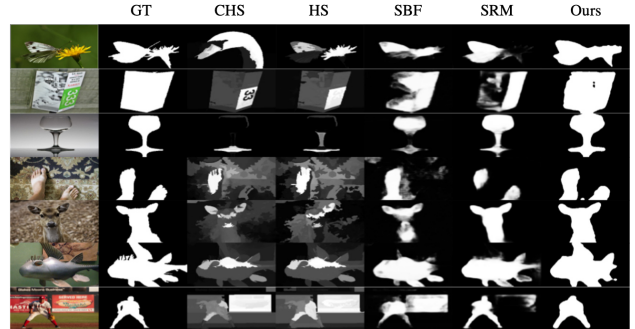


Figure 10. Visual results on the saliency prediction benchmarks. CHS [73], HS [84] (unsupervised), SBF [66] (unsupervised learning-based), SR [65] (supervised).

the static models on the DAVIS validation set using mIoU. Further, we perform the same evaluation on both FBMS and SegTrackV2, and report the scores in Tab. 2. As shown, with the confidence-aware adaptive bootstrapping loss Eq. (7), the static object model $\chi$ consistently improves over its counterpart on the three benchmarks, confirming the importance of uncertainty estimation in self-supervised learning.

|  | DAVIS | FBMS | SegTV2 |
|---|---|---|---|
| $\chi(F_\alpha)$ | 73.8 | 65.5 | 65.7 |
| $\chi(\mathcal{L}_\chi)$ | **78.2** | **68.7** | **69.3** |

Table 2. Static model $\chi(F_\alpha)$ vs. $\chi(\mathcal{L}_\chi)$. The latter is trained with confidence-aware adaptation.

### 5.3. Improvement from Bootstrapping

In Tab. 5, we show the improvement after multiple rounds of bootstrapping for each model as described in Algorithm 1 (dynamic-static bootstrapping). The reported score is the mIoU of the segmentation on the three video object segmentation benchmarks. This demonstrates the effectiveness of training in a synergistic loop with our dynamic-static bootstrapping strategy.

### 5.4. 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 $\{s_k\}_{k=1}^{2i+1}$, 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. 6, we report the performance of $\chi_i$'s, measured in terms of mIoU (with standard deviation computed across five runs). As shown in Tab. 6, when the number of the observed videos increases, the performance of the static object model also improves, which is consistently observed across multiple runs.

### 5.5. Video Object Segmentation Benchmarks

To check the effectiveness of the proposed dynamic-static bootstrapping in learning better motion segmentation, we evaluate on the benchmarks for video object segmentation. Since video object segmentation focuses on moving objects, we weigh the predictions from the dynamic model with the predictions from the static model to emphasize the detection of moving ones. Similar to [76], CRF postprocessing is performed to get our final results. We compare to top-performing unsupervised (no annotations are involved) and supervised (annotations are involved during the training phase) methods. The performance is measured by mIoU. As shown in Tab. 3, 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 2000 annotations. We have also listed the number of annotations used by other supervised methods in Tab. 3. Again, our model achieves the top performance using the least amount of manual annotations among all the supervised methods. In Tab. 7 we show the results by testing our trained models on two other video object segmentation benchmarks DAVIS17 and Youtube-VOS. Our model still achieves competitive performance compared to the state of the art and demonstrates good generalization.

### 5.6. Unsupervised Salient Object Detection

In Tab. 4, we evaluate the learned object prior (static object model) through the task of salient object detection in images. Note that the top-performing methods on unsupervised salient object detection all rely on handcrafted methods

| | Supervised Methods | | | | | Unsupervised Methods | | |
|---|---|---|---|---|---|---|---|---|
| | # Annot. | DAVIS | FBMS(J/F) | SegTV2 | | DAVIS | FBMS | SegTV2 |
| MATNet [81] | 14,000 | 82.4 | **76.1** / —— | 50.2 | ARP [30] | 76.2 | 59.8 | 57.2 |
| AnDiff [77] | 2,000 | 81.7 | —— / 81.2 | 48.3 | ELM [33] | 61.8 | 61.6 | – |
| COSNet [42] | 17,000 | 80.5 | 75.6 / —— | 49.7 | FST [48] | 55.8 | 47.7 | 47.8 |
| EPONet [13] | 2,000+ | 80.6 | – | 70.9 | NLC [14] | 55.1 | 51.5 | 67.2 |
| PDB [58] | 17,000 | 77.2 | 74.0 / 81.5 | 60.9 | SAGE [67] | 42.6 | 61.2 | 57.6 |
| LVO [63] | 2,000+ | 75.9 | 65.1 / 77.8 | 57.3 | STP [22] | 77.6 | 60.8 | 70.1 |
| FSEG [25] | 10,500 | 70.7 | 68.4 / —— | 61.4 | CIS [76] | 71.5 | 63.6 | 62.0 |
| Ours | 2,000 | **82.8** | 75.8 / **82.0** | **74.2** | Ours | **80.0** | **73.2** | **74.2** |

Table 3. Quantitative comparison on video object segmentation benchmarks with both supervised and fully unsupervised methods.

| | Supervised Methods | | | Unsupervised Methods | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *DSS* [19] | *NDF* [43] | *SR* [65] | RBD [82] | DSR [36] | HS [84] | CHS [66] | SBF [73] | USD [79] | DUSPS [46] | Ours |
| ECCSD | 87.9 | 89.1 | 82.6 | 65.2 | 63.9 | 62.3 | 68.2 | 78.7 | 87.8 | 87.4 | **88.1** |
| MSRA-B | 89.4 | 89.7 | 85.1 | 75.1 | 72.3 | 71.3 | 79.8 | – | 87.7 | **90.3** | 89.7 |
| DUT | 72.9 | 73.6 | 67.2 | 51.0 | 55.8 | 52.1 | – | 58.3 | 71.6 | 73.6 | **73.9** |

Table 4. Quantitative results on saliency prediction (or salient object detection) benchmarks.

| # of Rounds | DAVIS | FBMS | SegTV2 |
|---|---|---|---|
| 1 | 73.8 | 65.5 | 65.7 |
| 2 | 79.2 | 71.7 | 73.1 |
| 3 | **80.0** | **73.2** | **74.2** |

Table 5. The mIoU improves over multiple rounds of dynamic-static bootstrapping.

| # of videos | 12 | 36 | 60 | 84 | 108 |
|---|---|---|---|---|---|
| mIoU | 48.3 | 52.4 | 54.9 | 58.8 | 61.4 |
| Std. Dev. | 1.99 | 2.37 | 2.30 | 1.17 | 1.56 |

Table 6. Static model $\chi$ improves over time as more videos are observed.

| | MATNet [81] | Andiff [77] | ARP [30] | CIS [76] | Ours |
|---|---|---|---|---|---|
| DAVIS17 | 58.6 | 57.8 | 50.2 | 53.1 | **58.9** |
| YTVOS | – | 46.1 | 28.7 | 15.6 | **47.2** |

Table 7. Results on DAVIS17 [51] and Youtube-VOS [72] datasets.

either as the primary 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 [46]. 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. 4), our method still achieves competitive performance with no explicit annotation.

# 6. Discussion

We have presented a method to learn how to segment objects in images that exploits temporal consistency in their motion, observed in training videos, to bootstrap 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. 5, is the object a person? Or the motorcycle they are riding? The union of the two? The helmet they are 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 to consider them one object. If, in later video, a human is seen without a motorcycle, they will be an independent object thereafter. Admittedly, our model does not capture the fact that a proper object model should segment instances and enable multiple memberships 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. 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] Thomas D Albright and Gene R Stoner. Visual motion perception. *Proceedings of the National Academy of Sciences*, 92(7):2433–2440, 1995.

[3] Alper Ayvaci and Stefano Soatto. Detachable object detection: Segmentation and depth ordering from short-baseline video. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):1942–1951, 2011.

[4] Sven Bambach, David Crandall, Linda Smith, and Chen Yu. Toddler-inspired visual object learning. In *Advances in neural information processing systems*, pages 1201–1210, 2018.

[5] Adam Bielski and Paolo Favaro. Emergence of object segmentation in perturbed generative models. In *Advances in Neural Information Processing Systems*, pages 7254–7264, 2019.

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

[7] 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.

[8] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In *Advances in Neural Information Processing Systems*, pages 12705–12716, 2019.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Muhammad Faisal, Ijaz Akhter, Mohsen Ali, and Richard Hartley. Exploiting geometric constraints on dense trajectories for motion saliency. *arXiv preprint arXiv:1909.13258*, 2019.

[14] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting.

[15] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

[16] Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving objects in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4083–4090, 2015.

[17] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.

[18] E Bruce Goldstein and James Brockmole. *Sensation and perception*. Cengage Learning, 2016.

[19] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017.

[20] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2007.

[21] Ping Hu, Gang Wang, Xiangfei Kong, Jason Kuen, and Yap-Peng Tan. Motion-guided cascaded refinement network for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1400–1409, 2018.

[22] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 786–802, 2018.

[23] Serdar Ince and Janusz Konrad. Occlusion-aware optical flow estimation. *IEEE Transactions on Image Processing*, 17(8):1443–1451, 2008.

[24] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[25] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2117–2126. IEEE, 2017.

[26] 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.

[27] 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.

[28] Dominik A Klein and Simone Frintrop. Center-surround divergence of feature statistics for salient object detection. In *2011 International Conference on Computer Vision*, pages 2214–2219. IEEE, 2011.

[29] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[30] 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.

[31] 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.

[32] Zihang Lai and Weidi Xie. Self-supervised learning for video correspondence flow. *arXiv preprint arXiv:1905.00875*, 2019.

[33] Dong Lao and Ganesh Sundaramoorthi. Extending layered models to 3d motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

[34] Alex Levinshtein, Adrian Stere, Kiriakos N Kutulakos, David J Fleet, Sven J Dickinson, and Kaleem Siddiqi. Turbopixels: Fast superpixels using geometric flows. *IEEE transactions on pattern analysis and machine intelligence*, 31(12):2290–2297, 2009.

[35] 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.

[36] 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.

[37] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014.

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[39] Feng Liu and Michael Gleicher. Region enhanced scale-invariant saliency detection. In *2006 IEEE International Conference on Multimedia and Expo*, pages 1477–1480. IEEE, 2006.

[40] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018.

[41] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2010.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990.

[60] 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.

[61] 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.

[62] Brian Taylor, Vasiliy Karasev, and Stefano Soatto. Causal video object segmentation from persistence of occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4268–4276, 2015.

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

[64] John K Tsotsos, Scan M Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1-2):507–545, 1995.

[65] Tiantian 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.

[66] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[67] 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.

[68] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. Ranet: Ranking attention network for fast video object segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 3978–3987, 2019.

[69] 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.

[70] Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, 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.

[71] 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.

[72] 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.

[73] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1155–1162, 2013.

[74] 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.

[75] Yanchao Yang, Yutong Chen, and Stefano Soatto. Learning to manipulate individual objects in an image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6558–6567, 2020.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

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

[82] 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.

[83] 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.

[84] 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.