Unsupervised Moving Object Detection via Contextual Information Separation

Yanchao Yang*
UCLA Vision Lab

Antonio Loquercio*
University of Zurich

Davide Scaramuzza
University of Zurich

Stefano Soatto
UCLA Vision Lab

Abstract

We propose an adversarial contextual model for detecting moving objects in images. A deep neural network is trained to predict the optical flow in a region using information from everywhere else but that region (context), while another network attempts to make such context as uninformative as possible. The result is a model where hypotheses naturally compete with no need for explicit regularization or hyper-parameter tuning. Although our method requires no supervision whatsoever, it outperforms several methods that are pre-trained on large annotated datasets. Our model can be thought of as a generalization of classical variational generative region-based segmentation, but in a way that avoids explicit regularization or solution of partial differential equations at run-time. We publicly release all our code and trained networks.[1]

1. Introduction

Consider Fig. 1: Even relatively simple objects, when moving in the scene, cause complex discontinuous changes in the image. Being able to rapidly detect independently moving objects in a wide variety of scenes from images is functional to the survival of animals and autonomous vehicles alike. We wish to endow artificial systems with similar capabilities, without the need to pre-condition or learn similar-looking backgrounds. This problem relates to motion segmentation, foreground/background separation, visual attention, video object segmentation as we discuss in Sect. 3. For now, we use the words "object" or "foreground" informally[2] to mean (possibly multiple) connected regions of the image domain, to be distinguished from their surrounding, which we call "background" or "context," according to *some* criterion.

Since objects exist in the scene, not in the image, a method to infer them from the latter rests on an operational definition based on measurable image correlates. We call moving objects regions of the image whose motion cannot be explained by that of their surroundings. In other words, the motion of the background is uninformative of the motion of the foreground and vice-versa. The "information separation" can be quantified by the information reduction rate (IRR) between the two as defined in Sect. 2. This naturally translates into an adversarial inference criterion that has close connections with classical variational region-based segmentation, but with a twist: Instead of learning a generative model of a region that explains the image *in that region* as well as possible, our approach yields a model that tries to explain it *as poorly as possible* using measurements from *everywhere else but* that region.

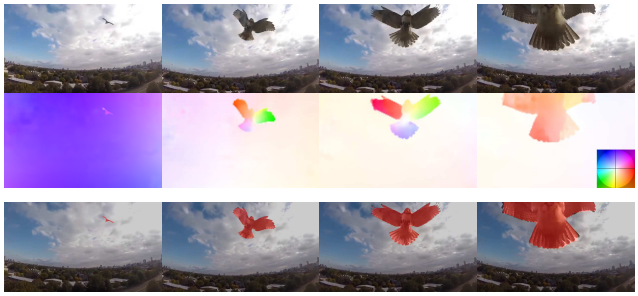

Figure 1: An encounter between a hawk and a drone (top). The latter will not survive without being aware of the attack. Detecting moving objects is crucial to the survival of animal and artificial systems alike. Note that the optical flow (middle row) is quite diverse within the region where the hawk projects: It changes both in space and time. Grouping this into a moving object (bottom row) is our goal in this work. Note the object is detected by our algorithm across multiple scales, partial occlusions from the viewpoint, and complex boundaries.

In generative model-based segmentation, one can always explain the image with a trivial model, the image itself. To avoid that, one has to impose model complexity bounds, bottlenecks or regularization. Our model does not have access to trivial solutions, as it is forced to predict a region without looking at it. What we learn instead is a contextual adversarial model, without the need for explicit regularization, where foreground and background hypotheses

[1] http://rpg.ifi.uzh.ch/unsupervised_detection.html

*These two authors contributed equally. Correspondence to yanchao.yang@cs.ucla.edu and loquercio@ifi.uzh.ch

[2] The precise meaning of these terms will be formalized in Sect. 2.

compete to explain the data with no pre-training nor (hyper)parameter selection. In this sense, our approach relates to adversarial learning and self-supervision as discussed in Sect. 3.

The result is a completely unsupervised method, unlike many recent approaches that are called unsupervised but still require supervised pre-training on massive labeled datasets and can perform poorly in contexts that are not well represented in the training set. Despite the complete lack of supervision, our method performs competitively even compared with those that use supervised pre-training (Sect. 4).

Summary of Contributions

Our method captures the desirable features of variational region-based segmentation: Robustness, lack of thresholds or tunable parameters, no need for training. However, it does not require solving a partial differential equation (PDE) at run-time, nor to pick regularizers or Lagrange multipliers, nor to restrict the model to one that is simple-enough to be tractable analytically. It also exploits the power of modern deep learning methods: It uses deep neural networks as the model class, optimizes it efficiently with stochastic gradient descent (SGD), and can be computed efficiently at run time. However, it requires no supervision whatsoever.

While our approach has close relations to both classical region-based variational segmentation and generative models, as well as modern deep learning-based self-supervision, discussed in detail in Sect. 3, to the best of our knowledge, it is the first *adversarial contextual model* to detect moving objects in images. It achieves better or similar performance compare to unsupervised methods on the three most common benchmarks, and it even edges out methods that rely on supervised pre-training, as described in Sect. 4. On one of the considered benchmarks, it outperforms all methods using supervision, which illustrates the generalizability of our approach. In Sect. 5 we describe typical failure modes and discuss limitations of our method.

2. Method

We call "moving object(s)" or "foreground" any region of an image whose motion is unexplainable from the context. A "region of an image" Ω is a compact and multiply-connected subset of the domain of the image, discretized into a lattice D. "Context" or "background" is the complement of the foreground in the image domain, $\Omega^c = D \backslash \Omega$. Given a measured image I and/or its optical flow to the next (or previous) image u, foreground and background are uncertain, and therefore treated as random variables. A random variable u_1 is "unexplainable" from (or "uninformed" by) another u_2 if their mutual information $\mathbb{I}(u_1; u_2)$ is zero, that is if their joint distribution equals the product of the marginals, $P(u_1, u_2) = P(u_1)P(u_2)$.

More specifically, the optical flow $u : D_1 \to \mathbb{R}^2$ maps the domain of an image $I_1 : D_1 \to \mathbb{R}^3_+$ onto the domain D_2 of I_2, so that if $x_i \in D_1$, then $x_i + u_i \in D_2$, where $u_i = u(x_i)$ up to a discretization into the lattice and cropping of the boundary. Ideally, if the brightness constancy constraint equation that defines optical flow was satisfied, we would have $I_1 = I_2 \circ u$ point-wise.

If we consider the flow at two locations i, j, we can formalize the notion of foreground as a region Ω that is uninformed by the background:

$$\begin{cases} \mathbb{I}(u_i, u_j | I) > 0, i, j \in \Omega \\ \mathbb{I}(u_i, u_j | I) = 0, i \in \Omega, j \in D \backslash \Omega. \end{cases} \tag{1}$$

As one would expect, based on this definition, if the domain of an object is included in another, then they inform each other (see appendix [40]).

2.1. Loss function

We now operationalize the definition of foreground into a criterion to infer it. We use the information reduction rate (IRR) γ, which takes two subsets $\boldsymbol{x}, \boldsymbol{y} \subset D$ as input and returns a non-negative scalar:

$$\gamma(\boldsymbol{x}|\boldsymbol{y}; I) = \frac{\mathbb{I}(u_{\boldsymbol{x}}, u_{\boldsymbol{y}}|I)}{\mathbb{H}(u_{\boldsymbol{x}}|I)} = 1 - \frac{\mathbb{H}(u_{\boldsymbol{x}}|u_{\boldsymbol{y}}, I)}{\mathbb{H}(u_{\boldsymbol{x}}|I)} \tag{2}$$

where $\mathbb{H}$ denotes (Shannon) entropy. It is zero when the two variables are independent, but the normalization prevents the trivial solution (empty set).[3] As proven in the appendix [40], objects as we defined them are the regions that minimize the following loss function

$$\mathcal{L}(\Omega; I) = \gamma(\Omega|\Omega^c; I) + \gamma(\Omega^c|\Omega; I). \tag{3}$$

Note that $\mathcal{L}$ *does not have a complexity term*, or regularizer, as one would expect in most region-based segmentation methods. This is a key strength of our approach, that involves no modeling hyperparameters, as we elaborate on in Sect. 3.

Tame as it may look, (3) is intractable in general. For simplicity we indicate the flow inside the region(s) Ω (foreground) with $u^{\text{in}} = \{u_i,\ i \in \Omega\}$, and similarly for u^{out}, the flow in the background Ω^c. The only term that matters in the IRR is the ratio $\mathbb{H}(u^{\text{in}}|u^{\text{out}}, I)/\mathbb{H}(u^{\text{in}}|I)$, which is

$$\frac{\int \log P(u^{\text{in}}|u^{\text{out}}, I) dP(u^{\text{in}}|u^{\text{out}}, I)}{\int \log P(u^{\text{in}}|I) dP(u^{\text{in}}|I)} \tag{4}$$

that measures the information transfer from the background to the foreground. This is minimized when knowledge of

[3] A small constant $0 < \epsilon \ll 1$ is added to the denominator to avoid singularities, and whenever $\boldsymbol{x} \neq \emptyset$, $\mathbb{H}(u_{\boldsymbol{x}}|I) \gg \epsilon$, thus we will omit ϵ from now on.

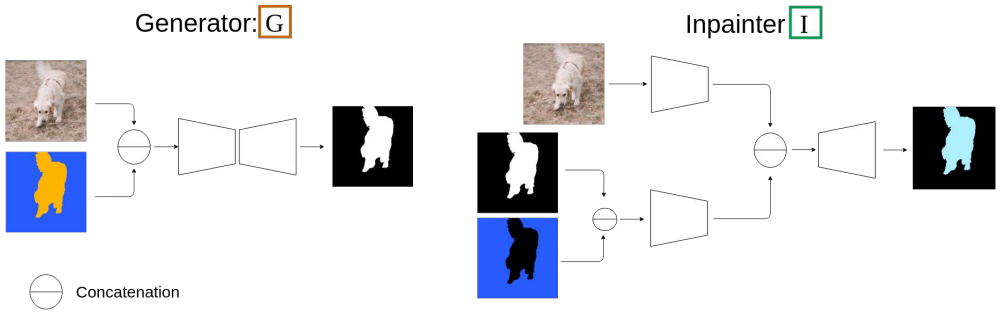

Figure 2: During training, our method entails two modules. One is the generator (G) which produces a mask of the object by looking at the image and the associated optical flow. The other module is the inpainter (I) which tries to inpaint back the optical flow masked out by the corresponding mask. Both modules employ the encoder-decoder structure with skip connections. However, the inpainter (I) is equipped with two separate encoding branches. See Sect. 4.1 for network details.

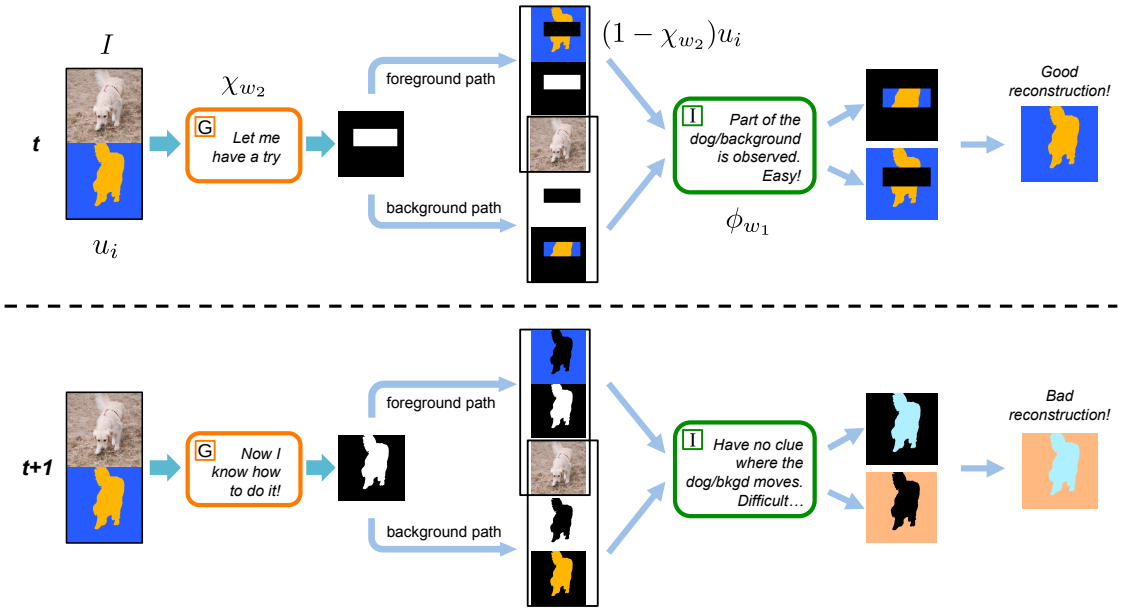

Figure 3: The two diagrams illustrate the learning process of the mask generator (G), after the inpainter (I) has learned how to accurately inpaint a masked flow. The upper diagram shows a poorly trained mask generator which does not precisely detect the object. Due to the imprecise detection, the inpainter can observe part of the object's flow, and perform an accurate reconstruction. At the same time, the inpainter partially observes the background's flow in the complementary mask. Consequently, it can precisely predict missing parts of the background's flow. In contrast, the lower diagram shows a fully trained mask generator which can precisely tell apart the object from the background. In this case, the inpainter observes the flow only outside the object and has no information to predict the flow inside it. At initialization time the inpainter does not know the conditionals to inpaint masked flows. Therefore, we propose to train both the generator and the inpainter jointly in an adversarial manner (see Sect. 2).

the background flow is sufficient to predict the foreground. To enable computation, we have to make draconian, yet common, assumptions on the underlying probability model, namely that

$$P(u^{\text{in}} = x|I) \propto \exp\left(-\frac{\|x\|^2}{\sigma^2}\right) \tag{5}$$

$$P(u^{\text{in}} = x|u^{\text{out}} = y, I) \propto \exp\left(-\frac{\|x - \phi(\Omega, y, I)\|^2}{\sigma^2}\right)$$

where $\phi(\Omega, y, I) = \int u^{\text{in}} dP(u^{\text{in}}|u^{\text{out}}, I)$ is the conditional mean given the image and the complementary observation. Here we assume $\phi(\Omega, \varnothing, I) = 0$, since given a single image the most probable guess of the flow is zeros. With these assumptions, (4) can be simplified, to

$$\frac{\int \|u^{\text{in}} - \phi(\Omega, u^{\text{out}}, I)\|^2 dP(u^{\text{in}}|u^{\text{out}}, I)}{\int \|u^{\text{in}}\|^2 dP(u^{\text{in}}|I)} \approx \approx \frac{\sum_{i=1}^{N} \|u_i{}^{\text{in}} - \phi(\Omega, u_i{}^{\text{out}}, I)\|^2}{\sum_{i=1}^{N} \|u_i{}^{\text{in}}\|^2} \tag{6}$$

where $N = |\mathcal{D}|$ is the cardinality of $\mathcal{D}$, or the number of flow samples available. Finally, our loss (3) to be minimized can be approximated as

$$\mathcal{L}(\Omega; I) = 1 - \frac{\sum_{i=1}^{N} \|{u_i}^{\text{in}} - \phi(\Omega, {u_i}^{\text{out}}, I)\|^2}{\sum_{i=1}^{N} \|{u_i}^{\text{in}}\|^2 + \epsilon} + 1 - \frac{\sum_{i=1}^{N} \|{u_i}^{\text{out}} - \phi(\Omega^c, {u_i}^{\text{in}}, I)\|^2}{\sum_{i=1}^{N} \|{u_i}^{\text{out}}\|^2 + \epsilon}. \quad (7)$$

In order to minimize this loss, we have to choose a representation for the unknown region Ω and for the function ϕ.

2.2. Function class

The region Ω that minimizes (7) belongs to the power set of D, that is the set of all possible subsets of the image domain, which has exponential complexity.[4] We represent it with the indicator function

$$\begin{aligned} \chi : D &\rightarrow \{0, 1\} \\ i &\mapsto 1 \text{ if } i \in \Omega;\ 0 \text{ otherwise} \end{aligned} \quad (8)$$

so that the flow inside the region Ω can be written as ${u_i}^{\text{in}} = \chi u_i$, and outside as ${u_i}^{\text{out}} = (1 - \chi)u_i$.

Similarly, the function ϕ is non-linear, non-local, and high-dimensional, as it has to predict the flow in a region of the image of varying size and shape, given the flow in a different region. In other words, ϕ has to capture the context of a region to *recover* its flow.

Characteristically for the ages, we choose both ϕ and χ to be in the parametric function class of deep convolutional neural networks, as shown in Fig. 2, the specifics of which are in Sect. 4.1. We indicate the parameters with w, and the corresponding functions ϕ_{w_1} and χ_{w_2}. Accordingly, after discarding the constants, the *negative* loss (7) can be written as a function of the parameters

$$\mathcal{L}(w_1, w_2; I) = \frac{\sum_i \|\chi_{w_2}(u_i - \phi_{w_1}(\chi_{w_2}, {u_i}^{\text{out}}, I))\|^2}{\sum_i \|{u_i}^{\text{in}}\|^2} + \frac{\sum_i \|(1 - \chi_{w_2})(u_i - \phi_{w_1}(1 - \chi_{w_2}, {u_i}^{\text{in}}, I)\|^2}{\sum_i \|{u_i}^{\text{out}}\|^2} \quad (9)$$

ϕ_{w_1} is called the *inpainter network*, and must be chosen to *minimize* the loss above. At the same time, the region Ω, represented by the parameters w_2 of its indicator function χ_{w_2} called *mask generator network*, should be chosen so that u^{out} is as uninformative as possible of u^{in}, and therefore the same loss is *maximized* with respect to w_2. This naturally gives rise to a minimax problem:

$$\hat{w} = \arg\min_{w_1} \max_{w_2} \mathcal{L}(w_1, w_2; I). \quad (10)$$

[4] In the continuum, it belongs to the infinite-dimensional set of compact and multiply-connected regions of the unit square.

This loss has interesting connections to classical region-based segmentation, but with a twist as we discuss next.

3. Related Work

To understand the relation of our approach to classical methods, consider the simplest model for region-based segmentation [8]

$$L(\Omega, c_i, c_o) = \int_{\Omega} |u^{\text{in}}(x) - c_i|^2 dx + \int_{\Omega^c} |u^{\text{out}}(x) - c_o|^2 dx \quad (11)$$

typically combined with a regularizing term, for instance the length of the boundary of Ω. This is a convex infinite-dimensional optimization problem that can be solved by numerically integrating a partial differential equation (PDE). The result enjoys significant robustness to noise, provided the underlying scene has piecewise constant radiance and is measured by image irradiance, to which it is related by a simple "signal-plus-noise" model. Not many scenes of interest have piecewise constant radiance, although this method has enjoyed a long career in medical image analysis. If we enrich the model by replacing the constants c_i with smooth functions, $\phi_i(x)$, we obtain the celebrated Mumford-Shah functional [25], also optimized by integrating a PDE. Since smooth functions are an infinite-dimensional space, regularization is needed, which opens the Pandora box of regularization criteria, not to mention hyperparameters: Too much regularization and details are missed; too little and the model gets stuck in noise-induced minima. A modern version of this program would replace $\phi(x)$ with a parametrized model $\phi_w(x)$, for instance a deep neural network with weights w pre-trained on a dataset $\mathcal{D}$. In this case, the loss is a function of w, with natural model complexity bounds. Evaluating ϕ_w at a point inside, $x \in \Omega$, requires knowledge of the entire function u *inside* Ω, which we indicate with $\phi_w(x, u^{\text{in}})$:

$$\int_{\Omega} |u^{\text{in}}(x) - \phi_w(x, u^{\text{in}})|^2 dx + \int_{\Omega^c} |u^{\text{out}}(x) - \phi_w(x, u^{\text{out}})|^2 dx. \quad (12)$$

Here, a network can just map $\phi_w(x, u^{\text{in}}) = u^{\text{in}}$ providing a trivial solution, avoided by introducing (architectural or information) bottlenecks, akin to explicit regularizers. We turn the table around and use the outside to predict the inside and vice-versa:

$$\int_{\Omega} |u^{\text{in}}(x) - \phi_w(x, u^{\text{out}})|^2 dx + \int_{\Omega^c} |u^{\text{out}}(x) - \phi_w(x, u^{\text{in}})|^2 dx \quad (13)$$

After normalization and discretization, this leads to our loss function (7). The two regions compete: for one to grow, the other has to shrink. In this sense, our approach relates to region competition methods, and specifically Motion Competition [12], but also to adversarial training, since we can

think of ϕ as the "discriminator" presented in a classification problem (GAN [1]), reflected in the loss function we use. This also relates to what is called "self-supervised learning," a misnomer since there is no supervision, just a loss function that does not involve externally annotated data. Several variants of our approach can be constructed by using different norms, or correspondingly different models for the joint and marginal distributions (5).

More broadly, the ability to detect independently moving objects is primal, so there is a long history of motion-based segmentation, or moving object detection. Early attempts to explicitly model occlusions include the layer model [38] with piecewise affine regions, with computational complexity improvements using graph-based methods [30] and variational inference [11, 6, 32, 43] to jointly optimize for motion estimation and segmentation; [26] use of long-term temporal consistency and color constancy, making however the optimization more difficult and sensitive to parameter choices. Similar ideas were applied to motion detection in crowds [5], traffic monitoring [4] and medical image analysis [14]. Our work also related to the literature on visual attention [16, 7].

More recent data-driven methods [36, 35, 9, 31] learn discriminative spatio-temporal features and differ mainly for the type of inputs and architectures. Inputs can be either image pairs [31, 9] or image plus dense optical flow [36, 35]. Architectures can be either time-independent [35], or with recurrent memory [36, 31]. Overall, those methods outperform traditional ones on benchmark datasets [26, 29], but at the cost of requiring a large amount of labeled training data and with evidence of poor generalization to previously unseen data.

It must be noted that, unlike in Machine Learning at large, it is customary in video object segmentation to call *"unsupervised"* methods that *do* rely on massive amounts of manually annotated data, so long as they do not require manual annotation at run-time. We adopt the broader use of the term where unsupervised means that there is no supervision of any kind both at training and test time.

Like classical variational methods, our approach does not need any annotated training data. However, like modern learning methods, our approach learns a contextual model, which would be impossible to engineer given the complexity of image formation and scene dynamics.

4. Experiments

We compare our approach to a set of state-of-the-art baselines on the task of video object segmentation to evaluate the accuracy of detection. We first present experiments on a controlled toy-example, where the assumptions of our model are perfectly satisfied. The aim of this experiment is to get a sense of the capabilities of the presented approach in ideal conditions. In the second set of experiments, we evaluate the effectiveness of the proposed model on three public, widely used datasets: Densely Annotated VIdeo Segmentation (DAVIS) [29], Freiburg-Berkeley Motion Segmentation (FBMS59) [26], and SegTrackV2 [37]. Provided the high degree of appearance and resolution differences between them, these datasets represent a challenging benchmark for any moving object segmentation method. While the DAVIS dataset has always a single object per scene, FBMS and SegTrackV2 scenes can contain multiple objects per frame. We show that our method not only outperforms the unsupervised approaches, but even edges out other supervised algorithms that, in contrast to ours, have access to a large amount of labeled data with precise manual segmentation at training time. For quantitative evaluation, we employ the most common metric for video object segmentation , *i.e.* the mean Jaccard score, a.k.a. intersection-over-union score, $\mathcal{J}$. Given space constraints, we add additional evaluation metrics in the appendix [40].

4.1. Implementation and Networks Details

Generator, *G*: Depicted on the left of Fig. 3, the generator architecture is a shrunk version of SegNet [2]. Its encoder part consists of 5 convolutional layers each followed by batch normalization, reducing the input image to $\frac{1}{4}$ of its original dimensions. The encoder is followed by a set of 4 atrous convolutions with increasing radius (2,4,8,16). The decoder part consists of 5 convolutional layers, that, with upsampling, generate an output with the same size of the input image. As in SegNet [2], a final softmax layer generates the probabilities for each pixel to be foreground or background. The generator input consists of an RGB image I_t and the optical flow $u_{t:t+\delta T}$ between I_t and $I_{t+\delta T}$, to introduce more variations in the optical flows conditioned on image I_t. At training time, δT is randomly sampled from the uniform distribution $\mathcal{U} = [-5, 5]$, with $\delta T \neq 0$. The optical flow $u_{t:t+\delta T}$ is generated with the pretrained PWC network [33], given its state-of-the-art accuracy and efficiency. The generator network has a total of 3.4M parameters.

Inpainter, *I*: We adapt the architecture of CPN [41] to build our inpainter network. Its structure is depicted on the right of Fig. 3. The input to this network consists of the input image I_t and the flow masked according to the generator output, χu, the latter concatenated with χ, to make the inpainter aware of the region to look for context. Differently from the CPN, these two branches are balanced, and have the same number of parameters. The encoded features are then concatenated and passed to the CPN decoder, that outputs an optical flow $\hat{u} = \phi(\chi, (1-\chi)u, I_t)$ of the same size of the input image, whose inside is going to be used for the difference between u^{in} and the recovered flow inside. Similarly, we can run the same procedure for the complement part. Our inpainter network has a total of 1.5M parameters.

At test time, only the generator G is used. Given I_t

	DAVIS [29]	FBMS59 [26]	SegTrackV2 [37]
$\mathcal{J} \uparrow$	92.5	88.5	92.1

Table 1: **Performance under ideal conditions:** When the assumptions made by our model are fully satisfied, our approach can successfully detect moving objects.. Indeed, our model reaches near maximum Jaccard score in all considered datasets.

and $u_{t:t+\delta T}$, it outputs a probability for each pixel to be foreground or background, $P_t(\delta T)$. To encourage temporal consistency, we compute the temporal average:

$$\overline{P_t} = \sum_{\delta T=-5, \neq 0}^{\delta T=5} P_t(\delta T) \tag{14}$$

The final mask χ is generated with a CRF [21] post-processing step on the final $\overline{P_t}$. More details about the post-processing can be found in the appendix.

4.2. Experiments in Ideal Conditions

Our method relies on basic, fundamental assumptions: *The optical flow of the foreground and of the background are independent*. To get a sense of the capabilities of our approach in ideal conditions, we artificially produce datasets where this assumption is fully satisfied. The datasets are generated as a modification of DAVIS2016 [29], FMBS [26], and SegTrackV2 [37]. While images are kept unchanged, ground truth masks are used to artificially perturb the optical flow generated by PWC [33] such that foreground and background are statistically independent. More specifically, a different (constant) optical flow field is sampled from a uniform distribution independently at each frame, and associated to the foreground and the background, respectively. More details about the generation of those datasets and the visual results can be found in the Appendix. As it is possible to observe in Table 1, our method reaches very high performance in all considered datasets. This confirms the validity of our algorithm and that our loss function (10) is a valid and tractable approximation of the functional (3).

4.3. Performance on Video Object Segmentation

As previously stated, we use the term *Unsupervised* with a different meaning with respect to its definition in literature of video object segmentation. In our definition and for what follows, the supervision refers to the algorithm's usage of ground truth object annotations at training time. In contrast, the literature usually defines methods as semi-supervised, if at test time they assume the ground-truth segmentation of the first frame to be known [3, 24]. This could be posed as tracking problem [42] since the detection of the target is human generated. Instead, here we focus on moving object detection and thus we compare our approach to the methods that are usually referred to as "unsupervised" in the video object segmentation domain. However we make further differentiation on whether the ground truth object segmentation is needed (supervised) or not (truly unsupervised) during training.

In this section we compare our method with other 8 methods that represent the state of the art for moving object segmentation. For comparison, we use the same metric defined above, which is the Jaccard score $\mathcal{J}$ between the real and predicted masks.

Table 2 shows the performance of our method and the baseline methods on three popular datasets, DAVIS2016 [29], FBMS59 [26] and SegTrackV2 [37]. Our approach is top-two in each of the considered datasets, and even outperforms baselines that need a large amount of labelled data at training time, *i.e.* FSEG [17].

As can be observed in Table 2, unsupervised baselines typically perform well in one dataset but significantly worse in others. For example, despite being the best performing unsupervised method on DAVIS2016, the performance of ARP [20] drops significantly in the FBMS59 [26] and SegTrackV2 [26] datasets. ARP outperforms our method by 6.5% on DAVIS, however, *our method outperforms ARP by 6.3% and 8.4%, on FBMS59 and SegTrackV2 respectively*. Similarly, NLC [15] and SAGE [39] are extremely competitive in the Segtrack and FBMS59 benchmarks, respectively, but not in others. NLC outperforms us on SegTrackV2 by 8.4%, however *we outperform NLC by 29.8% and 24.7%, on DAVIS and FBMS respectively*.

It has been established that being second-best in multiple benchmarks is more indicative of robust performance than being best in one [27]. Indeed, existing unsupervised approaches for moving object segmentation are typically highly-engineered pipeline methods which are tuned on one dataset but do not necessarily generalize to others. Also, consisting of several computationally intensive steps, extant unsupervised methods are generally orders of magnitude slower than our method (Table 3).

Interestingly, a similar pattern is observable for supervised methods. And this is particularly evident on the SegTrackV2 dataset [37], which is particularly challenging since several frames have very low resolution and are motion blurred. Indeed, supervised methods have difficulties with the covariate shift due to changes in the distribution between training and testing data. Generally, supervised methods alleviate this problem by pre-training on image segmentation datasets, but this solution clearly does not scale to every possible case. In contrast, our method can be finetuned on any data without the need for the latter to be annotated. As a result, our approach outperforms the majority of unsupervised methods as well as all the supervised ones, in

	PDB [31]	FSEG [17]	LVO [36]	ARP [20]	FTS [28]	NLC [15]	SAGE [39]	CUT [18]	Ours
DAVIS2016 [29] $\mathcal{J}\uparrow$	**77.2**	70.7	75.9	**76.2**	55.8	55.1	42.6	55.2	71.5
FBMS59 [26] $\mathcal{J}\uparrow$	**74.0**	68.4	65.1	59.8	47.7	51.5	61.2	57.2	**63.6**
SegTrackV2 [37] $\mathcal{J}\uparrow$	60.9	61.4	57.3	57.2	47.8	**67.2**	57.6	54.3	62.0
DNN-Based	Yes	Yes	Yes	No	No	No	No	No	Yes
Pre-Training Required	Yes	Yes	Yes	No	No	No	No	No	No

Table 2: **Moving Object Segmentation Benchmarks:** We compare our approach with 8 different baselines on the task of moving object segmentation. In order to do so, we use three popular datasets, *i.e.* DAVIS2016 [29], FBMS59 [26], and SegTrackV2 [37]. Methods in blue require ground truth annotations at training time and are pre-trained on image segmentation datasets. In contrast, methods in red are unsupervised and not require any ground-truth annotation. Our approach is top-two in all the considered benchmarks, comparing to the other unsupervised methods. **Bold** indicates best among all methods, while **Bold Red** and red represent the best and second best for unsupervised methods, respectively.

terms of segmentation quality and training efficiency.

4.4. Qualitative experiments and Failure Cases

In Fig. 4 we show a qualitative comparison of the detection generated by our and others' methods on the DAVIS dataset. Our algorithm can segment precisely the moving object regardless of cluttered background, occlusions, or large depth discontinuities. The typical failure case of our method is the detection of objects whose motion is due to the primary object. An example is given in the last row of Fig. 4, where the water moved by the surfer is also classified as foreground by our algorithm.

4.5. Training and Runtime Analysis

The generator and inpainter network's parameters are trained at the same time by minimizing the functional (10). The optimization time is approximately 6 hours on a single GPU Nvidia Titan XP 1080i. Since both our generator and inpainter networks are relatively small, we can afford very fast training/finetuning times. This stands in contrast to larger modules, *e.g.* PDB [31], that require up to 40 hrs of training.

At test time, predictions $\overline{P_t}$ (defined in eq. 14) are generated at 3.15 FPS, or with an average time of 320ms per frame, including the time to compute optical flow with PWC [33]. Excluding the time to generate optical flow, our model can generate predictions at 10.2 FPS, or 98ms per frame. All previous timings do not include the CRF post-processing step. Table 3 compares the inference time of our method with respect to other unsupervised methods. Since our method at test time requires only a pass through a relatively shallow network, it is orders of magnitude faster than other unsupervised approaches.

5. Discussion

Our definition of objects and the resulting inference criterion are related to generative model-based segmentation and region-based methods popular in the nineties. However, there is an important difference: Instead of using the evidence inside a region to infer a model of that region which is as accurate as possible, we use evidence *everywhere else but* that region to infer a model within the region, and we seek the model to be as bad as possible. This relation, explored in detail in Sect. 3, forces learning a contextual model of the image, which is not otherwise the outcome of a generative model in region-based segmentation. For instance, if we choose a rich enough model class, we can trivially model the appearance of an object inside an image region as the image itself. This is not an option in our model: We can only predict the inside of a region by looking outside of it. This frees us from having to impose modeling assumptions to avoid trivial solutions, but requires a much richer class of function to harvest contextual information.

This naturally gives rise to an adversarial (min-max) optimization: An inpainter network, as a discriminator, tries to hallucinate the flow inside from the outside, with the reconstruction error as a quality measure of the generator network, which tries to force the inpainter network to do the lousiest possible job.

The strengths of our approach relate to its ability to learn complex relations between foreground and background without any annotation. This is made possible by using modern deep neural network architectures like SegNet [2] and CPN [41] as function approximators.

Not using ground-truth annotations can be seen as a strength but also a limitation: If massive datasets are available, why not use them? In part because even massive is not large enough: We have shown that models trained on large amount of data still suffer performance drops whenever tested on a new benchmark significantly different from the training ones. Moreover, our method does not require any pre-training on large image segmentation datasets, and it can adapt to any new data, since it does not require any supervision. This adaptation ability is not only important for computer vision tasks, but can also benefit other applications, e.g. robotic navigation [23, 13] or manipulation [19].

Another limitation of our approach is that, for the task of

	ARP [20]	FTS [28]	NLC [15]	SAGE [39]	CUT [18]	Ours
Runtime(s)	74.5	0.5	11.0	0.88	103.0	**0.098**
DNN-based	No	No	No	No	No	Yes

Table 3: **Run-time analysis**: Our method is not only effective (top-two in each considered dataset), but also orders of magnitude faster than other unsupervised methods. All timings are indicated without optical flow computation.

Figure 4: **Qualitative Results:** We qualitatively compare the performance of our approach with several state-of-the-art baselines as well as the Ground-Truth (GT) mask. Our prediction are robust to background clutter, large depth discontinuities and occlusions. The last row shows a typical failure case of our method, *i.e.* objects which are moved by the primary objects are detected as foreground (water is moved by the surfer in this case).

motion-based segmentation, we require the optical flow between subsequent frames. One could argue that optical flow is costly, local, and error-prone. However, our method is general and could be applied to other statistics than optical flow. Such extensions are part of our future work agenda. In addition, our approach does not fully exploit the intensity image, although we use it as a conditioning factor for the inpainter network. An optical flow or an image can be ambiguous in some cases, but the combination of the two is rarely insufficient for recognition [43]. Again, our framework allows in theory exploitation of both, and in future work we intend to expand in this direction.

Acknowledgments

Yanchao Yang and Stefano Soatto would like to thank the support from ARO W911NF-17-1-0304 and ONR N00014-17-1-2072. Antonio Loquercio and Davide Scaramuzza are supported by the Swiss National Center of Competence Research Robotics (NCCR), through the Swiss National Science Foundation, and the SNSF-ERC starting grant.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 5

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image

segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(12):2481–2495, 2017. 5, 7

[3] L. Bao, B. Wu, and W. Liu. Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[4] D. Beymer, P. McLauchlan, B. Coifman, and J. Malik. A real-time computer vision system for measuring traffic parameters. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997. 5

[5] G. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2006. 5

[6] T. Brox, A. Bruhn, and J. Weickert. Variational motion segmentation with level sets. In *IEEE European Conference on Computer Vision (ECCV)*, pages 471–483. 2006. 5

[7] Z. Bylinskii, E. M. DeGennaro, R. Rajalingham, H. Ruda, J. Zhang, and J. K. Tsotsos. Towards the quantitative evaluation of visual attention models. *Vision research*, 116:258–268, 2015. 5

[8] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 10(2), 2001. 4

[9] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang. SegFlow: Joint learning for video object segmentation and optical flow. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 5

[10] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang. SegFlow: Joint learning for video object segmentation and optical flow. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 8

[11] D. Cremers and S. Soatto. Motion competition: A variational approach to piecewise parametric motion segmentation. *IEEE International Journal of Computer Vision*, 62(3):249–265, 2004. 5

[12] D. Cremers and S. Soatto. Motion competition: A variational approach to piecewise parametric motion segmentation. *International Journal of Computer Vision*, 62(3):249–265, 2005. 4

[13] P. Drews, G. Williams, B. Goldfain, E. A. Theodorou, and J. M. Rehg. Aggressive deep driving: Combining convolutional neural networks and model predictive control. In *Conference on Robot Learning (CoRL)*, 2017. 7

[14] A. Elnakib, G. Gimelfarb, J. S. Suri, and A. El-Baz. Medical image segmentation: a brief survey. In *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*, pages 1–39. Springer, 2011. 5

[15] A. Faktor and M. Irani. Video object segmentation by non-local consensus voting. In *British Machine Vision Conference (BMVC)*. British Machine Vision Association, 2014. 6, 7, 8

[16] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000. 5

[17] S. D. Jain, B. Xiong, and K. Grauman. FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6, 7

[18] M. Keuper, B. Andres, and T. Brox. Motion trajectory segmentation via minimum cost multicuts. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 7, 8

[19] O. Khatib. Real-time obstacle avoidance for manipulators and mobile robots. In *Proceedings. 1985 IEEE International Conference on Robotics and Automation*, volume 2, pages 500–505. IEEE, 1985. 7

[20] Y. J. Koh and C.-S. Kim. Primary object segmentation in videos based on region augmentation and reduction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6, 7, 8

[21] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. 6

[22] D. Lao and G. Sundaramoorthi. Extending layered models to 3d motion. In *IEEE European Conference on Computer Vision (ECCV)*, 2018. 8

[23] A. Loquercio, A. I. Maqueda, C. R. D. Blanco, and D. Scaramuzza. Dronet: Learning to fly by driving. *IEEE Robotics and Automation Letters*, 3(2):1088–1095, 2018. 7

[24] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. Video object segmentation without temporal information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 6

[25] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42(5):577–685, 1989. 4

[26] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(6):1187–1200, 2014. 5, 6, 7

[27] Y. Pang and H. Ling. Finding the best from the second bests-inhibiting subjective bias in evaluation of visual tracking algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2784–2791, 2013. 6

[28] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 7, 8

[29] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 6, 7

[30] J. Shi and J. Malik. Motion segmentation and tracking using normalized cuts. In *IEEE International Conference on Computer Vision (ICCV)*, 1998. 5

[31] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam. Pyramid dilated deeper ConvLSTM for video salient object detection. In *IEEE European Conference on Computer Vision (ECCV)*, 2018. 5, 7, 8

[32] D. Sun, J. Wulff, E. B. Sudderth, H. Pfister, and M. J. Black. A fully-connected layered model of foreground and background flow. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 5

[33] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5, 6, 7

[34] B. Taylor, V. Karasev, and S. Soatto. Causal video object segmentation from persistence of occlusions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015. 8

[35] P. Tokmakov, K. Alahari, and C. Schmid. Learning motion patterns in videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5, 8

[36] P. Tokmakov, K. Alahari, and C. Schmid. Learning video object segmentation with visual memory. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 5, 7

[37] D. Tsai, M. Flagg, and J. Rehg. Motion coherent tracking with multi-label MRF optimization. In *British Machine Vision Conference (BMVC) 2010*. British Machine Vision Association, 2010. 5, 6, 7

[38] J. Wang and E. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(5):625–638, 1994. 5

[39] W. Wang, J. Shen, and F. Porikli. Saliency-aware geodesic video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6, 7, 8

[40] Y. Yang, A. Loquercio, D. Scaramuzza, and S. Soatto. Unsupervised moving object detection via contextual information separation. *arXiv preprint arXiv:1901.03360*, 2019. 2, 5

[41] Y. Yang and S. Soatto. Conditional prior networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 271–287, 2018. 5, 7

[42] Y. Yang and G. Sundaramoorthi. Shape tracking with occlusions via coarse-to-fine region-based sobolev descent. *IEEE transactions on pattern analysis and machine intelligence*, 37(5):1053–1066, 2015. 6

[43] Y. Yang, G. Sundaramoorthi, and S. Soatto. Self-occlusions and disocclusions in causal video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4408–4416, 2015. 5, 8