International Journal of Computer Vision (2024) 132:40–55
https://doi.org/10.1007/s11263-023-01859-x

The Right Spin: Learning Object Motion from Rotation-Compensated Flow Fields

Pia Bideau[1] · Erik Learned-Miller[2] · Cordelia Schmid[3] · Karteek Alahari[4]

Received: 28 February 2022 / Accepted: 19 July 2023 / Published online: 18 August 2023

Abstract

A good understanding of geometrical concepts as well as a broad familiarity with objects lead to excellent human perception of moving objects. The human ability to detect and segment moving objects works in the presence of multiple objects, complex background geometry, motion of the observer and even camouflage. How we perceive moving objects so reliably is a longstanding research question in computer vision and borrows findings from related areas such as psychology, cognitive science and physics. One approach to the problem is to teach a deep network to model all of these effects. This is in contrast with the strategy used by human vision, where cognitive processes and body design are tightly coupled and each is responsible for certain aspects of correctly identifying moving objects. Similarly, from the computer vision perspective there is evidence that classical, geometry-based techniques are better suited to the "motion-based" parts of the problem, while deep networks are more suitable for modeling appearance. In this work, we argue that the coupling of camera rotation and camera translation can create complex motion fields that are difficult for a deep network to untangle directly. We present a novel probabilistic model to estimate the camera's rotation given the motion field. We then rectify the flow field to obtain a rotation-compensated motion field for subsequent segmentation. This strategy of first estimating camera motion, and then allowing a network to learn the remaining parts of the problem, yields improved results on the widely used DAVIS benchmark as well as the more recent motion segmentation data set MoCA (Moving Camouflaged Animals).

Keywords Motion segmentation · Video segmentation · Optical flow · Camera motion estimation

1 Introduction

The human visual system has the ability to detect independently moving objects in a large variety of different environments. While we are moving through the world our eye captures a large amount of visual information over time. Often, we are unaware of the remarkable preprocessing steps that happen almost unnoticed. For example, human eye movements induce two major simplifications to incoming images before visual information is processed by the visual cortex. These are: (1) stabilizing the image, i.e., reducing the amount of local change due to motion, and (2) changing the direction of gaze (Walls, 1962; Longuet-Higgins and Prazdny, 1980).

Here, we revisit this approach to motion segmentation and separate the problem into two parts: first, we preprocess the perceived motion field following well known geometrical concepts. This leads to important simplifications similar to gaze stabilization. Second, we learn to segment independently moving objects from these simplified motion fields (Fig. 1).

Prior work (Bideau and Learned-Miller, 2016a) provided a detailed overview and a general definition of the motion segmentation problem in computer vision. *To summarize, the task of motion segmentation attempts to analyze the perceived motion and to segment a video sequence into static environment (if any) and independently moving objects*. Interpreting the motion field accurately, and then drawing the right

Communicated by Matej Kristan.

✉ Pia Bideau
p.bideau@tu-berlin.de

[1] Technical University of Berlin, Marchstr, Berlin 10587, Germany

[2] University of Massachusetts, Governors Dr, Amherst, MA 01002, USA

[3] Inria, École Normale Supérieure, CNRS, PSL Research University, Paris, France

[4] Inria, CNRS, Grenoble INP, LJK, University Grenoble Alpes, 38000 Grenoble, France

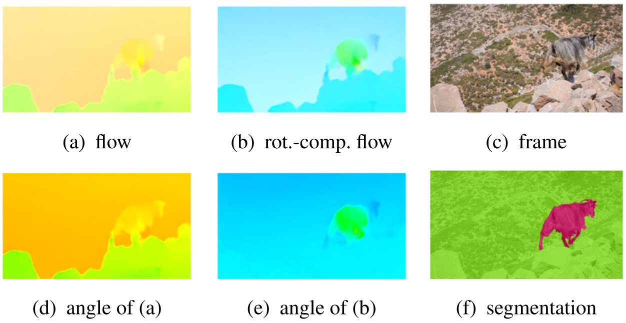

(a) flow (b) rot.-comp. flow (c) frame (d) angle of (a) (e) angle of (b) (f) segmentation

Fig. 1 What is moving? Coupling of camera rotation and camera translation often create complex motion fields that are difficult for a network to untangle. Instead we propose a strategy to learn object motion patterns based on rotation compensated flow

conclusions about what is moving in the world and what is static, is a complex process for synthetic systems. And even in biological vision systems applied strategies are still only partially understood. While motion segmentation refers to the task of segmenting objects based on their exhibited motion - and in particular their unique distinction from their static environment, we would like to emphasize the difference to the more general task of video object segmentation (VOS). Video object segmentation aims at segmenting a particular object throughout the video irrespective of its motion (Xu et al., 2018; Ding et al., 2023). The object to be segmented is typically the most dominant object or predefined in the first video frame. Due to the need of modeling spatio-temporal dependencies for video processing, video object segmentation is often influenced by insights obtained in motion segmentation.

Unlike most end-to-end learning-based approaches, where a model learns all the necessary steps between the input and the final output, we break down the problem of motion segmentation into two sub-problems: adjusting the optical flow to remove the effects of camera rotation (*rotation compensation*) using classical approaches based on perspective projection and learning to segment the remaining optical flow into static background and moving objects. The step of compensating for camera rotation is a challenging one, since the flow field is only a noisy estimate of the motion field. In cases of little motion or featureless areas, the observed flow field is often erroneous and thus the true camera motion and object motion is hard to estimate accurately. Prior work (Bideau et al., 2018; Bideau and Learned-Miller, 2016b) has explicitly addressed such challenges by incorporating a noise model into the camera motion estimation step. The approach proposed here builds upon the work of Bideau et al. (2018) and further refines a flow likelihood function that incorporates not only a model for the flow's noise, but also a new model for scene depth. To this end, we present here a novel probabilistic method for estimating camera rotation and derive a new likelihood function modeling the probability of an observed optical flow field, given an estimated (ideal) motion field. A CNN framework is then integrated for learning to segment moving objects after the motion of the camera has been determined.

Our contributions include: (i) estimating the camera rotation and translational motion direction in the presence of moving objects, using a new likelihood maximization approach, (ii) given the rotation compensated flow, we show that the task of learning motion patterns is improved, resulting in better motion segmentation performance shown on two data sets: the widely used DAVIS benchmark (Perazzi et al., 2016) and the recently published MoCA (Lamdouar et al., 2020). The latter focuses on the segmentation of camouflaged animals that are (close to being) invisible if they are not in motion.

The remainder of the paper is structured as follows. In Sect. 2 we review relevant work on motion segmentation starting from classical geometry-based approaches and concluding with the most recent ones using convolutional neural networks to segment moving objects from optical flow. In Sect. 3, we develop an end-to-end approach for motion segmentation. We briefly review the basics about the motion field and how it is related to camera motion, depth and object motion (Sect. 3.1). Building upon key concepts of perspective projection, the methodological approach is derived in two sections: estimating the camera rotation to produce rotation-compensated flow fields (Sect. 3.2) and segmenting the remaining (noisy) translational flow field into independently moving objects and static background (Sect. 3.3). A multifaceted evaluation of the proposed approach, including multiple ablation studies has been carried out and is shown in Sect. 4.

2 Related Work

Many works tackling the problem of motion segmentation focus on *binary motion segmentation*, where pixels are classified as either moving or being part of the static background. In that case no distinction is made between differently moving objects (Bideau and Learned-Miller, 2016b; Narayana et al., 2013; Papazoglou and Ferrari, 2013; Faktor and Irani, 2014). Others (Taylor et al., 2015; Keuper et al., 2015; Fragkiadaki et al., 2012) address *multi-label motion segmentation*, where a separate label is given to each independently moving object. Our work addresses binary motion segmentation, but we consider both views of the segmentation problem in this review of related work. We conclude the review of related work with a discussion of potential application areas: video object segmentation and the anticipation of object movement - two prominent lines of current research within the area of robotics and computer vision.

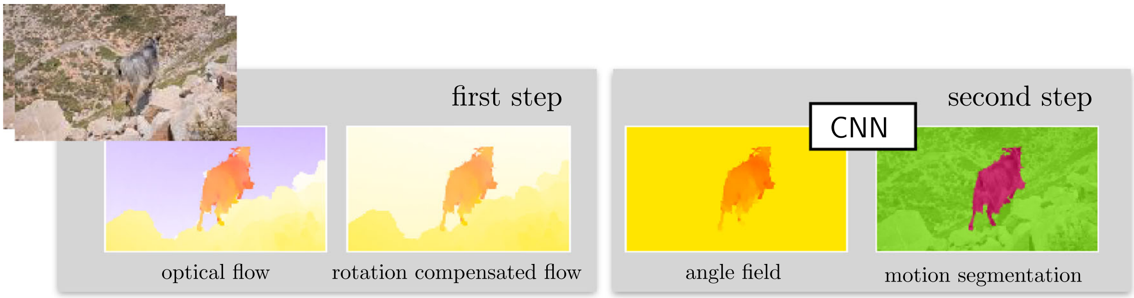


Fig. 2 Getting the right spin. We first compensate the observed motion field for camera rotation (“first step”), and segment the remaining translational optical flow field using a learning based approach (“second step”). The observed flow field on the left has complex motion patterns: the motion directions of foreground and background are pointing in opposite directions, due to large variance in scene depth, and the combined impact of camera rotation and translation. Estimating the camera rotation (“the right spin”), and compensating the flow field for this rotation simplifies the motion field dramatically, in this case yielding similar motion directions for foreground and background. This provides simpler inputs to our learning based motion segmentation framework

2.1 Classical Approaches

2.1.1 Methods Based on Feature Clustering

To capture motion information, typically point trajectories are either formed by tracked image features or dense optical flow. Then trajectories sharing similar motion characteristics are grouped into coherent motion clusters describing the motion of a particular object (Keuper et al., 2015; Brox and Malik, 2010; Fragkiadaki et al., 2012; Ochs and Brox, 2011; Keuper, 2017; Yan and Pollefeys, 2006; Shen et al., 2018; Lezama et al., 2011).

These approaches vary in defining typical motion characteristics for clustering. Yan and Pollefeys (2006) propose to cluster trajectories based on geometric constraints (trajectories of the same motion lie in a manifold) and locality. Keuper et al. (2015) represent the segmentation problem as a minimum cost multicut graph problem, where edge weights are computed from motion, position and color cues.

These trajectory based clustering approaches reach their limit if understanding of the scene structure is necessary to segment a moving object correctly. Trajectories perfectly represent long-term pixel displacements between a sequence of frames. Pixel displacements however are a function of depth and motion. Thus trajectory based clustering methods often form clusters not only for independently moving objects, but also for objects at different depths. For instance if the camera is translating and rotating rocks close to the camera produce a very different flow pattern that the far away scene (see Fig. 2), thus those two areas would form two separate clusters although neither the rock nor the far away scene is moving.

Methods based on occlusions (Ogale et al., 2005; Taylor et al., 2015) are subject to similar depth-related problems, since occlusions could be caused at depth boundaries as well as motion boundaries. A distinction is often not made.

2.1.2 Methods Based on Projective Geometry

Euclidean geometry describes well the three-dimensional world. Under Euclidean transformations (rotation and translation) certain geometric properties of the world do not change - lengths, angles between intersecting lines, parallel lines stay parallel. While euclidean geometry describes the three-dimensional world well, it is insufficient to describe the imaging process of the camera. Here, lengths and angles are no longer preserved and parallel lines actually may meet at a distance. To this end many computer vision methods rely on projective geometry, an extension of the euclidean geometry (Torr, 1998; Zamalieva and Yilmaz, 2014; Wang and Adelson, 1994; Ke and Kanade, 2002; Jin et al., 2008; Xiao and Shah, 2005; Vidal and Ma, 2004; Xu et al., 2018). Projective geometry models a much larger class of transformations than just rotations and translations. Projective geometry also includes important transformations such as perspective projections among many others. This has the advantage of being able to model the imaging process of the camera, but on the other hand comes with drawbacks as fewer measures are preserved - lengths, angles and parallelism. Preserving fewer measures allows modeling the imaging process of the camera but simultaneously reducing measures in this case allows unrealistic deformations such as shearing.

Different from trajectory based clustering methods, motion segmentation approaches relying on projective geometry analyze the optical flow between a pair of frames, grouping pixels into regions where flow is consistent with motion models that are explainable by projective geometry (Torr, 1998; Zamalieva and Yilmaz, 2014; Wang and Adelson, 1994; Ke and Kanade, 2002; Jin et al., 2008; Xiao and Shah, 2005; Xu et al., 2018). Torr (1998) develops a sophisticated probabilistic model of optical flow, building a mixture model that explains an arbitrary number of rigid components within the scene. Interestingly, he assigns different types of motion models to each object based on model fitting criteria. Zamalieva and Yilmaz (2014) and Xu et al. (2018) present a combination of methods that rely on both - projective geometry (homography estimation) and perspective projection (fundamental matrix estimation). The two methods have complimentary strengths, and the authors attempt to select among the best dynamically.

Methods relying on projective geometry perform well in cases of planar motion (motion obtained by a translating or rotating camera picturing a planar scene or a very distant scene, where effects of 3D parallax are negligible), however similarly to cluster based approaches these methods fall short in case of complex scene geometry.

Horn identified specific drawbacks of using projective geometry in such estimation problems and has argued that methods based directly on perspective projection are less prone to overfitting in the presence of noise (Horn, 1999) as those come with fewer and physically plausible invariants.

2.1.3 Methods Based on Perspective Projection

Perspective geometry allows us to mathematically explain and model the process of how the three-dimensional world is projected on to a two-dimensional image plane. Artists and scientists like Alberti, Brunelleschi, Dürer and da Vinci studied effects of perspective projection about 500 years back in time (Pirenne, 1952). These insights have made a significant contribution to current successes in computer vision. One of the key aspects of perspective projection is the observation that two parallel lines (in the euclidean space) are transformed to two lines that intersect in the vanishing point at the horizon on the image plane. Perspective projection transformations are one out of many transformations allowed in projective geometry.

It has been shown that motion segmentation approaches based on perspective projection (Irani and Anandan, 1998; Bideau and Learned-Miller, 2016b; Bideau et al., 2018; Narayana et al., 2013; Vidal et al., 2002; Zhang et al., 2007; Yang and Ramanan, 2021) are more accurate (in terms of model agreement to the physical world) than those based on projective geometry, since the latter omits certain constraints in modeling image transformations (Horn, 1999; Bideau and Learned-Miller, 2016b). Having a model that conforms to the physical world might be especially critical for tasks where the focus lies on real world interaction such as in robotics and autonomous driving scenarios.

2.2 Learning motion segmentation using convolutional neural networks

Methods based on supervised learning Several approaches in computer vision have explored the strength of deep neural networks to learn motion patterns of moving objects and to produce binary motion masks distinguishing whether a pixel belongs to a moving object or not (Tokmakov et al., 2017a, b; Jain et al., 2017; Cheng et al., 2017; Dave et al., 2019; Ranjan et al., 2019; Vertens et al., 2017; Mahadevan et al., 2020; Lamdouar et al., 2020; Cheng et al., 2017). Most approaches propose a two-stream architecture to separately process motion and appearance (Tokmakov et al., 2017b; Jain et al., 2017; Dave et al., 2019).

Theses approaches learn motion patterns given the optical flow, the raw video frames or optical flow together video frames. Rather than following the true physics of image formation, convolutional neural networks are able to learn high level motion patterns of background motion and object motion. This ability has the clear advantage of not being dependent upon technical camera parameters such as the focal length or image distortions due to various lens characteristics or constraints induced by technical parts of the camera (mechanical or electronic).

2.2.1 Methods Based on Self-supervised Learning

General concerns of deep-learning based approaches and in particular supervised approaches are overfitting to a particular type of object category that is likely to move (Dave et al., 2019) and the lack of large amounts of training data. To overcome the problem of limited training data, two straight forward approaches are either using synthetic training data (Tokmakov et al., 2017a, b) or relying on noisy estimates of the motion field (Jain et al., 2017) using other algorithms (Sun et al., 2018; Ilg et al., 2017; Sun et al., 2010). However, both paths are still in need of large amounts of training data (although no additional manual annotations are required in these cases), this rises the need for self-supervised approaches (Yang et al., 2021; Lu et al., 2019; Yang et al., 2019; Lai et al., 2020; Gordon et al., 2019; Bideau et al., 2018). Incorporating knowledge about the real world physics into the training procedure of a neural network is an alternative to various kinds of data augmentation approaches that is subject of current research (Tung et al., 2019; Yang and Ramanan, 2021). Some of those ideas have been already successfully applied in context of self-supervised learning (Gordon et al., 2019; Bideau et al., 2018).

In this work, we propose a novel approach to the motion segmentation problem that specifically combines aspects of perspective projection and learns general object motion patterns.

2.3 Application Areas for Motion Segmentation

Motion segmentation relies on low-level visual properties such as optical flow and thus is object agnostic. This key property has not only been discovered for semi-supervised video object segmentation (Luiten et al., 2019), it has also

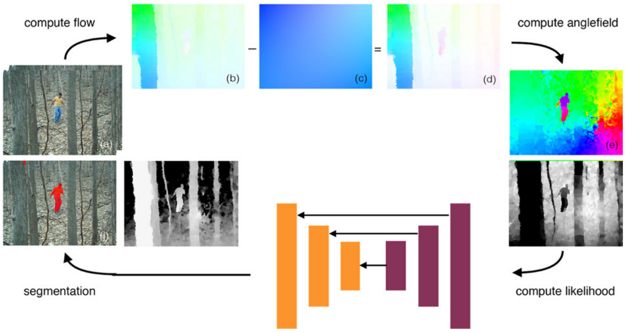


Fig. 3 Overview of our approach. Given the optical flow **b** the camera rotation is estimated (Sect. 3.2.4). The flow $\mathbf{v_r}$ due to camera rotation is defined by the motion parameters (A, B, C). **c** is subtracted from the optical flow **o** to produce a translational flow $\mathbf{o}_t$. The flow angle $\theta_{\mathbf{o}_t}$ and magnitudes $|\mathbf{o}_t|$ are shown in **e**

enabled a new line of research dealing with object segmentation by anticipating motion (Choudhury et al., 2022). Here, an image segmentation network is supervised by a simple pretext task of predicting image regions that are likely to move leading to an unsupervised approach for video and image segmentation by anticipating motion. Today, understanding the movement of objects has significantly influenced several lines of research areas - video object segmentation (VOS) and the anticipation of object movement may be the most prominent areas. Latter shows high potential for key research questions for robotics - manipulation and robot-environment interaction, where a prior guess about object movement is essential (Eisner et al., 2022; Xu et al., 2022).

3 Learning Object Motion from Rotation-Compensated Flow

Like many previous works, we define a *moving object* as a *collection of matter that independently moves as a whole in the 3D world*. An overview of our approach for motion segmentation is shown in Fig. 3. Given an estimate of the motion field (optical flow) each frame is segmented into static environment and independently moving objects. To achieve this we present an approach where we first estimate the camera rotation and then use this knowledge to form a rotation-compensated flow field. A network is trained that takes rotation-compensated flow fields as input and outputs motion segmentation masks. To this end, we combine our novel geometry-based method for estimating camera rotation, and a CNN framework for learning to segment moving objects.

In the following we will revise relevant background information about the formation of a motion field, that occurs on the camera sensor as the camera moves (Sect. 3.1). Building on this, we propose a novel approach to estimate camera rotation in complex environments, considering scene depth as well as independently moving objects (Sect. 3.2). In Sect. 3.3, we propose an approach similar to Bideau et al. (2018) that learns to segment the rotation compensated motion field into static background and independently moving objects.

3.1 The Motion Field: A Geometrical Analysis

The motion field captures pixel displacements between two consecutive frames. Displacements arise typically due to one of the following factors: (1) a moving camera, (2) one or more objects moving in the 3D world. These pixel displacements depend not only on the speed of objects or the camera, but also the scene geometry.

As an example to illustrate the different factors that influence the formation of the motion field, let's consider the "goat" sequence from the DAVIS data set (Fig. 2). Based on the original flow field it is hard to estimate which pixels belong to the moving object and which belong to static background. The direction of the flow in the background region differs significantly from the flow describing the motion of the rocks in the foreground region (motion direction is color encoded). However, neither the background nor the rocks are moving differently in the 3D world. To detect objects that are actually moving independently in 3D it is necessary to decompose the observed motion field. We formalize these observations and review the geometrical construction of the motion field.

3.1.1 Motion Field

Let $[U, V, W]$ be the parameters describing the camera translation and $[A, B, C]$ the parameters describing camera rotation[1] along the x, y and z axes respectively. Let f be the camera's focal length and Z the relative scene depth at a pixel location (x, y). In this setting, the motion vector $\mathbf{v}$ due to camera motion is given by:

$$\mathbf{v} = \mathbf{v}_r + \mathbf{v}_t = \begin{pmatrix} u_r \\ v_r \end{pmatrix} + \begin{pmatrix} u_t \\ v_t \end{pmatrix}, \tag{1}$$

$$= \begin{pmatrix} \frac{A}{f}xy - Bf - \frac{B}{f}x^2 + Cy \\ Af + \frac{A}{f}y^2 - \frac{B}{f}xy - Cx \end{pmatrix} + \begin{pmatrix} \frac{-fU+xW}{Z} \\ \frac{-fV+yW}{Z} \end{pmatrix}, \tag{2}$$

where $\mathbf{v}_r$ and $\mathbf{v}_t$ represent motion field vectors corresponding to camera rotation and translation respectively. Equation (2).[2]

[1] The rotation parameters are often referred to as pitch, yaw and roll.

[2] In fact, this equation is an approximation, and only holds if the rotation angles are small (Longuet-Higgins and Prazdny, 1980) To obtain the exact rotational flow field one has to transform the 2D image points to 3D using perspective projection equations, rotate the points according to the camera's rotation in 3D, backproject them onto the 2D image plane, and then measure the displacement.

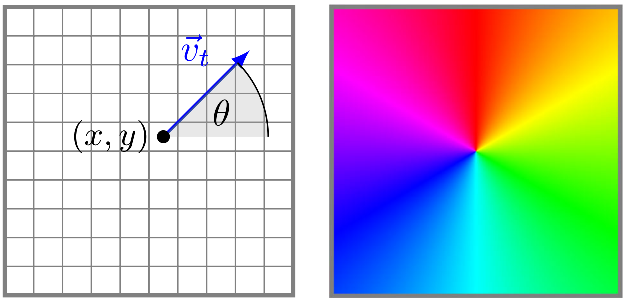


Fig. 4 Translational motion field vector. Left: motion field vector $\mathbf{v}_t$ at pixel position (x, y). Right: color coding of the angle field $\theta(x, y)$ at each pixel location for the case of camera translation along the optical axis, i.e. $[U, V, W] = [0, 0, 1]$

highlights an important properly, namely that the flow due to camera rotation is only determined by the camera rotation parameters and the camera's focal length. *The flow due to camera rotation is independent of the scene depth.* One can subtract this rotational motion component at each pixel to obtain a rotation-compensated flow field.

3.1.2 Rotation-Compensated Motion Field

As shown in the flow Eq. (2), the rotation-compensated flow field $\mathbf{v}_t$ is determined by the translational camera motion $[U, V, W]$, and the scene depth Z. It comprises all the relevant information about the scene geometry, unlike the rotational component $\mathbf{v}_r$, which is independent of the scene geometry. The magnitude of the rotation-compensated flow is inversely related to scene depth, i.e., regions further away from the camera have small translational flow magnitude, and those closer to the camera have larger magnitudes. *The direction of* $\mathbf{v_t}$ *(flow angle) however does not depend upon the scene depth*:

$$\theta = \begin{cases} \arccos\,(xW - fU), & \text{if } (yW - fV) > 0, \\ 2\pi - \arccos\,(xW - fU), & \text{otherwise.} \end{cases} \quad (3)$$

Figure 4 pictures the computation of the flow angle θ at pixel locations (x, y), leading to an angle field as shown on the right. Where as Fig. 4 pictures the angle field of pure camera translation, Fig. 2 shows an angle field of a scene with camera translation and object motion. Here, independently moving objects, can be observed as discontinuities in angle. The angle of the rotation-compensated flow alone is independent of the scene geometry, thus independently moving objects stand out due to their different direction.

3.2 The Right Spin: Camera Motion Estimation

To rectify the observed optical flow field for camera rotation, we require an accurate estimate for rotation. How can we obtain a good estimate of the camera rotation and the translational motion direction that together best explain the observed motion field? Towards finding an answer to this question, we derive a novel maximum likelihood approach that aims at finding the rotation $[A, B, C]$ such that the likelihood of the resulting translational flow field is maximized. To this end, we derive a new flow likelihood function incorporating a model for the optical flow's noise as well as a prior distribution over the *inverse scene depth*.

In the following, we first introduce the new flow likelihood (Sect. 3.2.1). We then describe how camera motion parameters are estimated by maximizing this new likelihood function.

3.2.1 Likelihood of the Translational Flow

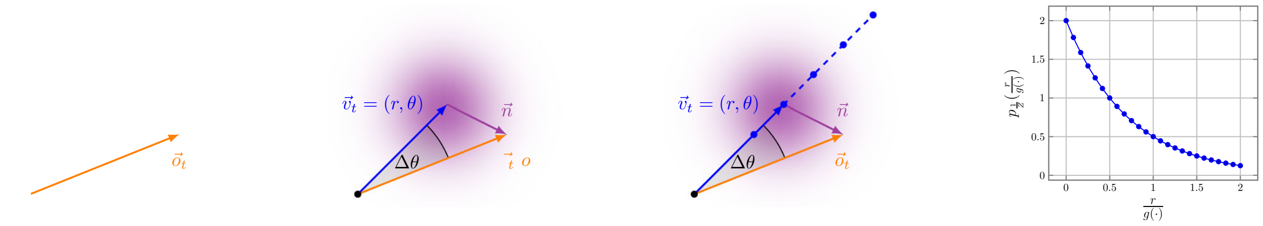
Let $\mathbf{o}_t$ be the observed translational flow vector, e.g., flow estimated with (Sun et al., 2018), at the pixel position (x, y). Let the translational 3D motion direction of the camera $[U, V, W]$ be a unit vector. The three translational camera parameters $[U, V, W]$ and the pixel position (x, y) define the direction of a motion field vector on the image plane. As derived in (Bideau et al., 2018), the probability of observing $\mathbf{o}_t$ at (x, y) given a motion direction $[U, V, W]$ is:

$$p(\mathbf{o}_t \mid U, V, W, x, y) = \int_0^\infty p(\mathbf{n})\ p(r \mid U, V, W, x, y)\, dr, \quad (4)$$

This likelihood function explicitly incorporates a model for the optical flow's noise - the distribution over the optical flow's noise $p(\mathbf{n})$, and a model for the motion field magnitude - the distribution over motion field magnitude $p(r \mid U, V, W, x, y)$. The variable r denotes the magnitude of a motion field vector and $\mathbf{n}$ denotes the optical flow's noise.

Modeling the probability distribution over flow magnitudes r is challenging, since those depend on the camera's translational motion direction $[U, V, W]$, the pixel location as well as the scene depth at that location. Prior work (Bideau et al., 2018) models the probability distribution over flow magnitudes by assuming that the motion field magnitude r is independent of $[U, V, W]$. However this may lead to inaccuracies, especially in the case of strong z-motion of the camera (forward motion). In this case motion field magnitudes close to the focus of expansion are near zero and the motion vectors farther away from the focus of expansion show larger magnitudes, thus the motion field magnitude is clearly dependent upon the camera's motion direction $[U, V, W]$.

A better approach than assuming independence is expressing the motion field magnitude as a function of the inverse scene depth, which is inversely proportional to the motion field magnitude r. Next, we present a new way of model-



(a) Translational (observed) flow vector $\vec{o}_t$ at pixel location (x, y).

(b) Observed optical flow vector $\vec{o}_t$, is a noisy observation of the motion field vector $\vec{v}_t$: $\vec{o}_t = \vec{v}_t + \vec{n}$.

(c) To compute the flow likelihood, we integrate over the unknown motion magnitude of the motion field vector $\vec{v}_t$.

(d) Probability distribution over inverse depth.

Fig. 5 Flow likelihood. **a–c** computation of the probability $p(\mathbf{n})$ at pixel location (x, y). **d** probability distribution over inverse depth. The flow likelihood is maximal, when the observed flow vector $\mathbf{o}_t$ and the motion field vector $\mathbf{v}_t$ point into the same direction with similar magnitude

ing the distribution over motion field magnitudes without the need of making independence assumptions.

3.2.2 Distributions Over Flow Magnitudes Expressed as a Function of Inverse Depth

We express the motion field magnitudes in terms of inverse depth $\frac{1}{Z}$ and $g(\cdot)$. The function $g(\cdot)$ comprises all aspects of the flow magnitude that are *not* related to depth,

$$\begin{aligned} r &= \sqrt{u_t^2 + v_t^2}, \\ &= \frac{1}{Z} \cdot g(f, x, y, U, V, W). \end{aligned} \tag{5}$$

Given this reformulation of the magnitude r, we can determine the *induced distribution* over motion field magnitudes, given the distribution over inverse depths. We aim to compute $p(r \mid g(f, x, y, U, V, W))$ through $p(\frac{1}{Z})$, which is the distribution over inverse depth. Using the relation between r and $g(\cdot)$ from Eq. (5), we can rewrite $p(r \mid g(\cdot))$ as follows

$$p(r \mid g(\cdot)) = \frac{p(\frac{1}{Z})}{g(\cdot)}. \tag{6}$$

This is effectively just a change of units. Expressing the distribution over flow magnitudes in terms of the distribution over inverse depth however brings a significant advantage. This formulation effectively factors motion direction (U, V, W), focal length f and scene depth into the function $g(\cdot)$, and the distribution over depth can be modeled without relying on these dependencies that require making further approximations.

3.2.3 New Flow Likelihood

Following prior derivation, the flow likelihood function (Eq. 4) can be defined by the distribution over the flow's noise $p(\mathbf{n})$ and the distribution over inverse depth, instead of flow magnitudes:

$$\begin{aligned} &p(\mathbf{o_t} \mid U, V, W, x, y) \\ &\quad = \int_0^\infty p(\mathbf{n}) \, \frac{p\left(\frac{1}{Z}\right)}{g(\cdot)} \, dZ. \end{aligned} \tag{7}$$

Figure 5 pictures the distribution of $p(\mathbf{n})$, with $\mathbf{n}$ being the noise added to the unknown motion field vector v_t leading to the observed flow o_t. The probability of the flow noise $p(\mathbf{n})$ is modeled as a multivariate normal distribution $p(\mathbf{n}) \sim \mathcal{N}(\mu, \Sigma)$. The inverse depth $p(\frac{1}{Z})$ is modeled as an exponential distribution $p(\frac{1}{Z}) \sim \text{Exp}(\lambda)$. Details regarding the parametrization of these two distributions can be found in Sect. 3.4.

3.2.4 Camera Motion Estimation via Likelihood Maximization

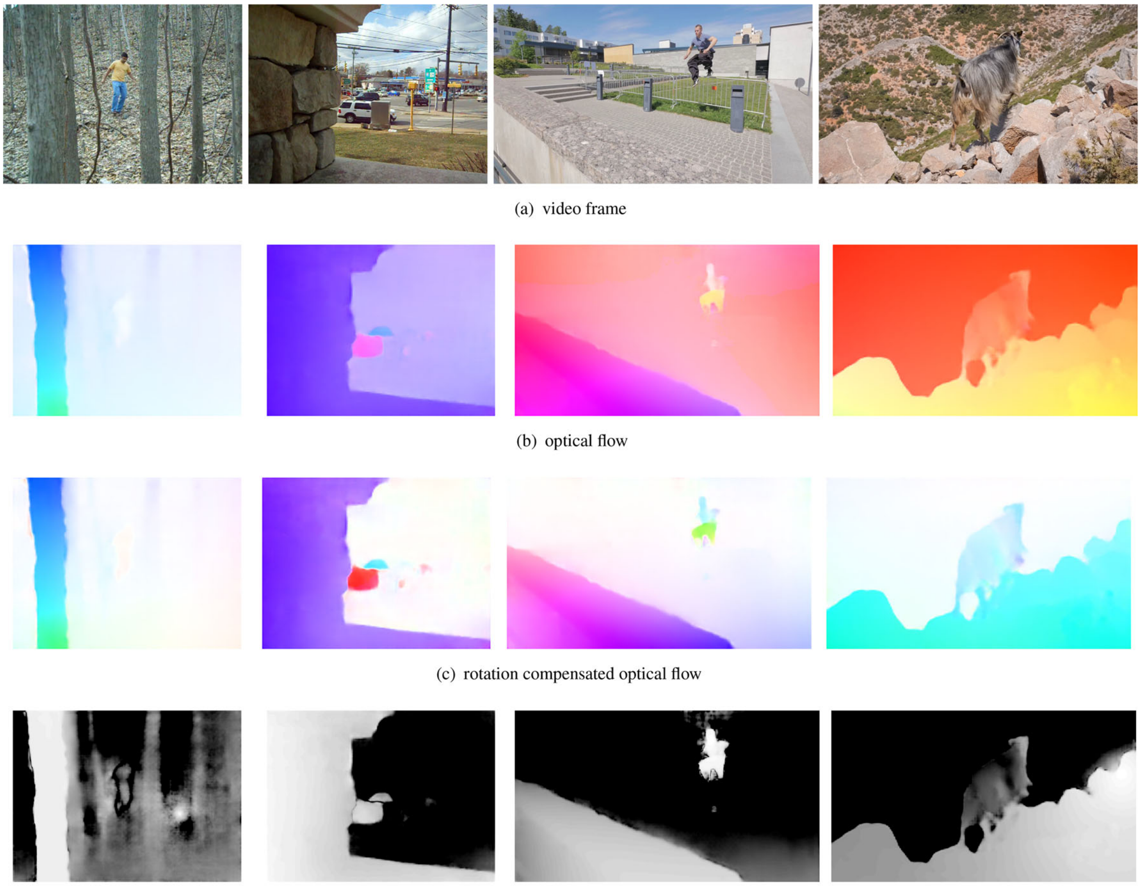
In Sect. 3.2.1, we have derived a new likelihood function of an observed optical flow vector $\mathbf{o}$. Our goal is now to find a camera rotation (A, B, C) and translational camera motion direction (U, V, W), such that the flow likelihood is maximal or alternatively the negative log-likelihood is minimal. Recall $\mathbf{o}_t$ is the observed translational flow vector after subtracting the flow $\mathbf{v}_r$ due to camera rotation:

$$\mathbf{o}_t = \mathbf{o} - \mathbf{v}_r(A, B, C). \tag{8}$$

Given the rotation compensated flow, we minimize the negative log-likelihood as follows:

$$\begin{aligned} &A^*, B^*, C^*, U^*, V^*, W^* \\ &= \underset{A,B,C,U,V,W}{\arg\min} \sum -\log(p(\mathbf{o_t} \mid U, V, W, x, y)). \end{aligned} \tag{9}$$

Local minima are a concern when solving this optimization problem, especially in cases of noisy optical flow, inaccurate estimates of independently moving objects or complex scene geometry. To reduce this risk, we initialize the optimization


(a) video frame

(b) optical flow

(c) rotation compensated optical flow

(d) depth estimate

Fig. 6 Flow, rotation compensated flow and the relative depth estimate. We show sample videos from the data set Complex Background (video sequences: traffic, forest) as well as two sample videos from the DAVIS data set (video sequence: parkour, goat). A comparison of **b** and **d** shows how motion at distant is dominated by camera rotation. After subtracting of the camera's rotation the remaining flow magnitude in these areas is very small (light color). If the flow magnitude is small the motion direction is noisy. This can be seen in (**e**)

using three different starting points: (1) camera rotation and translation estimate of the previous frame, (2) camera rotation estimate weighted by depth estimate of the previous frame and the translation estimate of the previous frame, and (3) camera rotation estimate weighted by depth estimate of the previous frame and the translation estimate of the previous frame in the opposite direction. The first initialization is a good assumption if the camera motion is approximately constant. Initialization (2) and (3) incorporate depth information. The apparent motion of areas far away is mainly influenced by the camera's rotation and not the camera's translation (see Fig. 6), thus knowing the depth helps to correctly disentangle flow due to camera rotation and flow due to translation.

During the optimization each pixel is weighted using learned, soft object motion masks of the previous frame, that evolve over time - thus the influence of moving objects is suppressed due to a low weight. The following Section describes how object motion masks are learned while pertaining important geometric information.

3.3 Object Motion Segmentation

We build our segmentation framework on an effective model for motion segmentation, that learns object motion patterns from optical flow and segments a flow field into static background and moving objects (Tokmakov et al., 2017a). Yet, this model does not incorporate any geometrical concepts. As discussed earlier optical flow fields couple information about scene geometry as well as camera motion, making the judgment whether an object is moving challenging. By introducing a simple pre-processing step we show, that the complexity of optical flow patterns is dramatically reduced. Different from prior work, our network processes rotation compensated flow fields (angle + magnitude) to segment independently moving objects. Learning object motion based on pre-processed flow fields appears to be an easier task to learn. While our network architecture is similar to (Tokmakov et al., 2017a), we propose important modifications to the training procedure in the following.

3.3.1 Incorporating Geometric Information into Training

The network follows the classical U-Net architecture and is trained on estimated translational flow fields. During training, we first estimate optical flow on the FlyingThings3D data set (Mayer et al., 2016) using the flow estimation algorithm by Sun et al. (2018). The ground truth camera rotation is provided and subtracted from the estimated flow to obtain a rotation-compensated flow field. This flow field is input to our network. The input has a size of $h \times w \times 3$. The third dimension denotes the flow expressed in terms of angle (represented as a unit vector) and magnitude. A representation of the flow angle as unit vector instead of angles in degree avoids segmentation discontinuities at zero degree (or 2π respectively). The normalized flow field and the flow's magnitude are concatenated and form the input to our network. An interesting question for training a network with rotation-compensated optical flow is, whether it is worthwhile to incorporate the magnitude into the training procedure. On the one hand the flow magnitude can be a good indicator about the reliability of the flow angle (Bideau and Learned-Miller, 2016b), while on the other hand variation in larger magnitudes can be either due to variances in the scene depth or fast moving objects - thus including the magnitude might add rather misleading information. We take a closer look into this question as part of our ablation study in Sect. 4.2.

3.4 Implementation Details

To find the camera rotation and translational motion direction that best explains the observed optical flow field, we derived a new flow likelihood function (Sect. 3.2). Details regarding parametrization are provided in the following.

The probability of the flow noise $p(\mathbf{n})$ is modeled as a multivariate normal distribution $p(\mathbf{n}) \sim \mathcal{N}(\mu, \Sigma)$ and the inverse depth $p(\frac{1}{Z})$ as an exponential distribution $p(\frac{1}{Z}) \sim \text{Exp}(\lambda)$. The noise covariance Σ is assumed to be spherical and is measured using the ground truth flow of Sintel (Butler et al., 2012) and the corresponding noisy estimate Sun et al. (2018). We obtain $\Sigma = 16.5 \cdot 10^{-5} I$, where I is the identity matrix. λ is the rate parameter of the exponential distribution modeling the inverse depth, and is estimated using ground truth depths from Sintel. We measured $\lambda = 0.64$. The distribution over inverse depth can be seen in Fig. 5d.

For computational efficiency the integral in Eq. (7) is approximated using a discrete sum over motion field magnitudes r. Flow likelihood values are pre-computed and stored in a lookup table for efficiency (see Fig. 7).

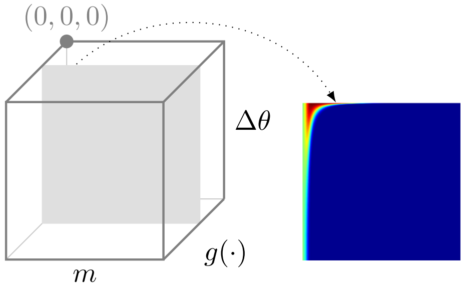


Fig. 7 Lookup table picturing flow likelihood values. Our new flow likelihood addresses the challenge of estimating the camera's motion in the presence of noisy optical flow. The color *red* indicates high likelihood values, *dark blue* indicates low likelihood values. The lower the angle difference $\Delta\theta$ between the vectors $\mathbf{o}_t$ and $\mathbf{v}_t$, the higher the likelihood. Note that for very small flow magnitudes m the flow likelihood is almost the same regardless $\Delta\theta$. This is an important consequence of our model, indicating the unreliability of the flow direction in case of near zero magnitudes

4 Experiments

We begin with a brief description of data sets used for training and evaluation of our motion segmentation network. In Section 4.1, we evaluate our here presented motion segmentation approach on the widely used DAVIS data set (Perazzi et al., 2016) and MoCA (Lamdouar et al., 2020). Ground truth camera motion is not provided for these data sets, thus synthetic data - such as the FlyingThings3D data set (Mayer et al., 2016) and Sintel (Butler et al., 2012; Wulff et al., 2012) - are used for ablation studies. These studies in particular focus on the analysis of different variants of our core network and the quality as well as the effect of rotation estimation via likelihood maximization.

DAVIS2016 (Densely Annotated VIdeo Segmentation) contains 50 video sequences in total with moving objects in various environments. A 30/20 training/validation split is provided. Our model is evaluated on the validation set. Ground truth segmentations of the most prominent moving object are provided for each frame. DAVIS has been widely used for general video segmentation as well as motion segmentation.

MoCA (Moving camouflaged animals) comprises a set of 141 videos depicting 67 different animals. The data set is split into three motion types describing the animals motion - *locomotion*, *static* and *deformation*. Following the procedure of (Lamdouar et al., 2021; Yang et al., 2019) we evaluate on the *locomotion* split, which forms the largest part of the dataset with 88 video sequences in total. Annotations are provided in form of bounding boxes. An evaluation script is provided by the authors of MoCA.

FT3D (FlyingThings3D) is a large optical flow data set, providing ground truth optical flow, RGB images, camera motion and depth. It is a synthetic data set showing random

Table 1 Motion segmentation: Comparison to approaches solely relying motion cues on DAVIS2016 ***(train-val)***

		LMP	TMM	Ours	Ours*
	Supervised	✓	✗	✓	✓
$\mathcal{J}$	Mean ↑	58.4	40.1	59.7	**62.5**
	Recall ↑	67.3	34.3	69.6	**73.8**
	Decay ↓	5.6	15.2	4.3	**3.8**
$\mathcal{F}$	Mean ↑	58.4	39.6	59.5	**61.1**
	Recall ↑	66.0	15.4	66.4	**69.9**
	Decay ↓	7.9	12.7	**5.4**	5.6
$\mathcal{T}$	Mean ↓	87.8	**51.3**	74.5	83.4

Ours refers to the variant of our model using only motion cues and no appearance terms and Ours* denotes a motion-only upper bound, which uses ground truth segmentation for camera motion estimation. Best viewed in color (**1st-best**, 2nd-best)

objects like chairs, tables, etc., flying in a 3D world along random trajectories. FT3D is split into test and training set.

Sintel is the de facto benchmark for optical flow algorithms, containing 23 video sequences with 20 to 50 frames each. These short video sequences are taken from an animated movie. Synthetic videos are available with ground truth optical flow, depth, camera motion and material segmentation.

4.1 Results

Our main framework consists of two steps (1) compensating the observed optical flow for camera rotation, and (2) segmenting the resulting optical flow in to static background and independently moving objects. Experiments presented here are based on the DAVIS data set and the MoCA data set, that each raise a slightly different aspect onto the motion segmentation problem. Details are described in the following.

DAVIS: Optical flow only We compare our motion segmentation network with other methods that use optical flow as the only cue for segmentation. Table 1 shows these results on DAVIS. LMP (Tokmakov et al., 2017a) is a learning based approach trained on ground truth optical flow of FlyingThings3D. This approach relies on a simliar network architecture, but does not incorporate an explicit model for modeling geometrical concepts, e.g. the scene geometry and camera motion. TMM (Bideau and Learned-Miller, 2016b), on the contrary, compensates flow for camera rotation and attempts to segment a video by assigning translational motion models to different image regions in a probabilistic fashion. The exclusive usage of translational motion models however quickly leads to oversegmentations and fails to capture more complex motion patterns. While combining geometrical concepts such as perspective projection together with learned motion patterns, our approach improves over both these motion segmentation methods. The segmentation performance is measured using the $\mathcal{J}$-Mean score. We achieve an $\mathcal{J}$-Mean score of 59.7. The next best performing method is LMP resulting in an $\mathcal{J}$-Mean score of 58.4. We compute an upper bound for our method (Ours* in Table 1) by masking out independently moving objects, with ground truth segments, for our camera motion estimation procedure. This masking procedure eliminates errors of our camera motion estimation due to 'outliers' in optical flow, such as moving objects.

MoCA

Data sets like MoCA focus in particular on the segmentation of objects that can only be robustly recognized based on their unique motion. Where as most data sets for moving object segmentation combine several cues (motion and appearance) that are helpful for recognizing moving objects, this data set highlights the relevance of motion. Thus MoCA allows to evaluate the strengths of motion models in isolation. It is not surprising that appearance cues are rather weak in cases of camouflage, therefore methods based on RGB frames only (e.g. COSNet by Lu et al. (2019)) show a weak performance in these settings (see Table 2). Similarly MATNet, which is typically jointly used with CRF post-processing, enhancing the segmentation quality of video frames along their corresponding RGB images, shows a significant performance loss. Our model capturing motion patterns in a temporal consistent manner over multiple frames (Ours-Temp) outperforms all supervised approaches on MoCA with respect to $\mathcal{J}$-Mean accuracy. The performance of MATNet (with multi-frame input) is comparable to ours if used without CRF post-processing.

Our approach taking a single optical flow frame (compensated for camera rotation) as input, performs comparable to other supervised approaches. A simple post-processing step - convolution with a 3D Gaussian filter and frame-wise application of a dense CRF, eliminates temporal instabilities (Ours+Temp in Table 2). Among all methods SegI (Lamdouar et al., 2021) shows best results on MoCA; on DAVIS their performance falls short due to their lack of a strong appearance model. SegI combines multiple ConvNets where each of them encode a flow frame together with a transformer network without taking RGB frames into consideration. The model is trained on synthetically generated data, and thus can be considered as unsupervised. In contrast, our approach was trained using rotation compensated flow frames estimated from the synthetic dataset *FlyingThings3D*.

DAVIS: Optical flow + Appearance Our main contribution lies in a novel approach for learning to segment moving objects based on optical flow only. We incorporate appearance information similar to LVO (Tokmakov et al., 2017a) and compare to segmentation approaches that consider both - appearance as well as motion information (Table 3). Within

Table 2 Motion segmentation: Comparison to state-of-the-art motion segmentation methods on MoCA

		SegI	MG	CIS	ARP	COD	COSNet	MATNet w/o crf	MATNet	Ours	Ours+Temp
	Supervised	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓
	RGB	✗	✗	✓	✓	✓	✓	✓	✓	✗	✗
	Flow	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
	Multi-frame	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓
$\mathcal{J}$	Mean ↑	**68.6**	63.4	49.4	61.2	44.9	50.7	64.7	54.9	58.3	**65.8**
Success Rate	$\tau = 0.5$	**77.2**	74.2	55.6	67.6	41.4	58.8	72.3	59.8	64.5	**72.7**
	$\tau = 0.6$	**71.7**	65.4	46.3	60.6	33.0	53.4	**67.1**	55.3	58.0	65.2
	$\tau = 0.7$	**62.3**	52.4	32.9	51.3	23.5	45.7	**59.8**	48.5	49.8	53.8
	$\tau = 0.8$	**46.4**	35.1	17.6	37.6	14.0	33.7	**46.4**	39.0	36.2	38.9
	$\tau = 0.9$	**25.5**	14.7	3.0	20.0	5.9	16.7	**22.1**	21.6	16.6	17.3
	SR_{mean}	**56.6**	48.4	31.1	47.4	23.6	41.7	**53.5**	44.8	45.0	49.6

Methods we compare against from left to right: (Lamdouar et al., 2021; Yang et al., 2021, 2019; Koh and Kim, 2017; Lamdouar et al., 2020; Lu et al., 2019; Zhou et al., 2020). Bold indicates best among all methods, while **1st-best** and 2nd-best represent the best and second best within the supervised methods. Best viewed in color

Table 3 Motion segmentation: Comparison to state-of-the-art motion segmentation methods on DAVIS2016

		SFL	COSNet	MATNet	LMP+App	FSEG	LVO	Ours+App	CIS	ARP	SegI
	Supervised	✤	✤	✤	✓	✓	✓	✓	✗	✗	✗
	RGB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
	Flow	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓
	Multi-frame	✗	✗	✓	✗	✗	✓	✓	✓	✓	✓
$\mathcal{J}$	Mean ↑	67.4	80.5	**82.4**	70.0	70.7	72.2	**73.5**	71.5	76.2	67.8
	Recall ↑	81.4	94.0	**94.5**	85.0	83.5	82.4	**85.5**	86.5	91.1	-
	Decay ↓	6.2	**0.0**	5.5	1.3	1.5	**0.1**	1.2	9.5	7.0	-
$\mathcal{F}$	Mean ↑	66.7	79.4	**80.7**	65.9	65.3	67.5	**68.9**	70.5	70.6	-
	Recall ↑	77.1	**90.4**	90.2	79.2	73.8	75.4	**79.6**	83.5	83.5	-
	Decay ↓	5.1	**0.0**	4.5	2.5	1.8	2.7	**1.4**	7.0	7.9	-

We group approaches according their training strategy: supervised and trained on the DAVIS training split (✤), supervised and trained on other segmentation data sets (✓) and unsupervised methods (✗). Methods we compare against from left to right: (Cheng et al., 2017; Lu et al., 2019; Zhou et al., 2020; Tokmakov et al., 2017a; Jain et al., 2017; Tokmakov et al., 2017b; Yang et al., 2019; Koh and Kim, 2017; Lamdouar et al., 2021; Yang et al., 2021). Bold indicates best among all methods, while **1st-best** and 2nd-best represent the best and second best within the supervised methods. Best viewed in color

the group of supervised approaches our approach shows the best performance in terms of mean/recall $\mathcal{J}$ and $\mathcal{F}$. Where as ours and LVO integrate appearance cues in a similar manner, these approaches differ in the way that object motion cues are learned. LVO learns object motion patterns directly from optical flow, whereas we first disentangle camera rotation and translation before segmenting independently objects. Ablation studies analyze the usefulness of this disentanglement in further detail. Within the unsupervised approaches ARP (Koh and Kim, 2017), which is a non learning based approach, reaches highest performance. Due to multiple iterations over the entire video this approach is computationally expensive as mentioned by Yang et al. (2021). Among all methods MATNet reaches highest accuracy in terms of mean $\mathcal{J}$ and $\mathcal{F}$. One reason might lie in their training strategy, which makes use of the DAVIS training set (indicated with ✤).

A qualitative comparison with the best performing methods is shown in Fig. 8. Our results based on optical flow only and based on optical flow in combination with appearance are shown in the last two rows of this figure. These two rows in particular highlight the complementarity of motion and appearance cues. We miss the hiker's foot when relying on motion alone (Ours), since it is not moving. However, while integrating motion with appearance, we segment the entire object accurately. ARP, the strongest method among unsupervised approaches, relies on segmenting the primary object(s) in a video and comes with a noticeable bias towards the object's appearance. In many cases such a strong appearance model is advantageous. However, it can lead to erroneous

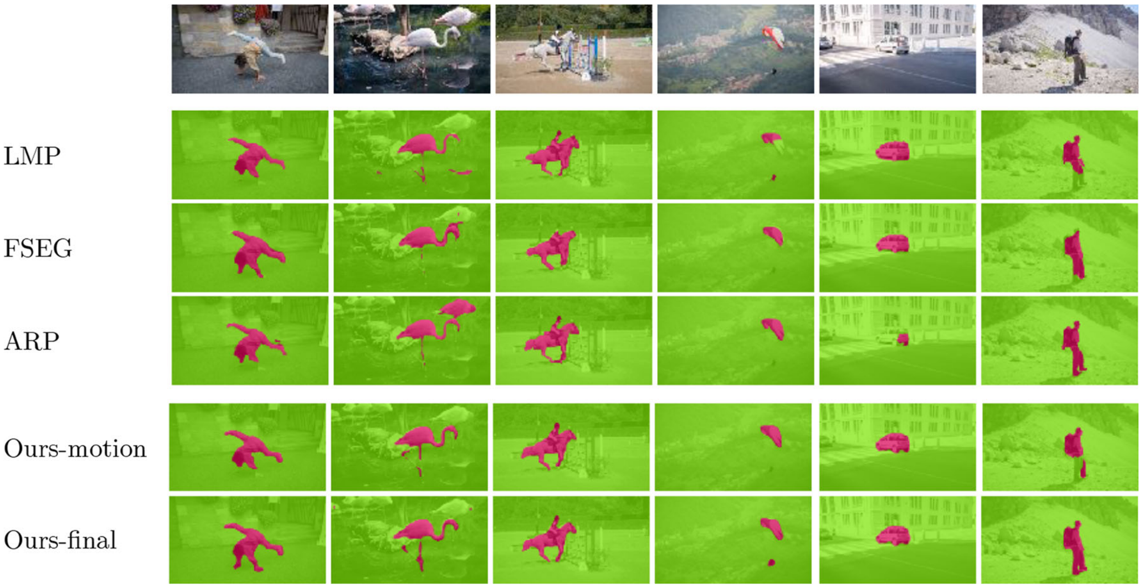


Fig. 8 Qualitative segmentation results. Qualitative segmentation results on the DAVIS data set, showing a comparison with three other best performing methods. Ours-final denotes our complete method and Ours the variant based on motion cues alone

segmentations in other cases. For example, it only segments a part of the car in Fig. 8, 2nd column from the right. The car moves from the darker (shadow) area to the brighter (sunny) region and is only partially segmented because only a portion of the object matches the primary object's appearance. Our method that extracts geometrical information from optical flow and integrates learned objectness cues is capable of overcoming these types of failure cases relying on appearance.

4.2 Ablation Study

Network variants - Effectiveness of rotation compensation We trained four variants of our motion segmentation network, with: (1) ground truth optical flow, (2) the ground truth flow after removing ground truth camera rotation, i.e., with rotation compensated-flow fields, (3) estimated optical flow field using PWC-Net (Sun et al., 2018), and (4) estimated ground truth flow compensated with ground truth camera rotation, i.e., estimated rotation compensated-flow field. Table 4 shows the analysis with these four variants. Training and testing with ground truth optical flow (original: gt flow or compensated: gt t-flow) leads to significantly better than estimation results than using estimated optical flow. Segmentation accuracy is about 20% higher on the FT3D test set for ground truth, compared to estimated optical flow. Training with rotation-compensated optical flow consistently leads to improved quality of the final segmentation, e.g., 90.68% vs. 93.23%, which supports the idea behind our method. While results on ground truth flow confirm the conceptual idea of facilitating the input flow field, experiments based on estimated optical flow (e.g. PWC flow) show realistically achievable results in case of noisy optical flow. We conclude, learning can be significantly simplified, if we are able to efficiently incorporate knowledge about physical concepts into the process of moving object segmentation. The benefit of incorporating knowledge about physical concepts in particular matters in case of complex scene geometry, where a coupling of camera rotation and translation leads to complex flow fields. A visual comparison in terms of segmentation quality between using the original flow as input instead of the rotation-compensated flow is shown in Fig. 9.

Table 4 Ablation study: Network variants - effectiveness of rotation compensation

Trained with..	Tested with..	Angle+magnitude
gt flow	gt flow	90.68
gt t-flow	gt t-flow	93.23
PWC flow	PWC flow	77.18
PWC t-flow	PWC t-flow	78.69

We trained four networks using flow angle and magnitude with: the provided ground truth optical flow of FT3D (Mayer et al., 2016) (gt flow), ground truth optical flow after subtracting ground truth camera rotation (gt t-flow), estimated optical flow using (Sun et al., 2018) (PWC flow), and estimated optical flow after subtracting ground truth camera rotation (PWC t-flow). Segmentation accuracy is measured on the FT3D test set with intersection over union (IoU) scores

Training on flow angle only versus angle+magnitude As discussed in Sect. 3.1, rotation-compensated flow comprises all the information about independent object motion and the scene structure (depth). In this context, two interesting questions to tackle are: *how well can one extract informa-*

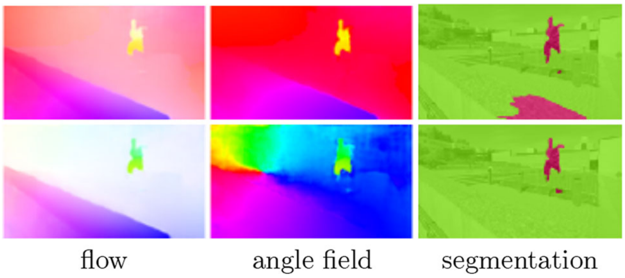


Fig. 9 Ablation study: Comparison of motion segmentation results based on the original and the rotation-compensated flow field. Top row: motion segmentation with the original flow field that includes camera rotation, translation and object motion. Bottom row: motion segmentation based on *rotation-compensated flow field*. Note that the angle field (middle) of the rotation-compensated flow is entirely depth independent. The angle field is fully determined by the translational camera motion and object motion. In this example one can observe a clear z-motion of the camera, which is shown by the rainbow pattern. The angle field of the original flow containing both camera rotation and translation is depth dependent (top row, middle image). This angle field clearly shows discontinuities in angle at the wall, which is due to significant changes in depth and not because of independent object motion

Table 5 Ablation study: Training with flow angle vs flow angle and magnitude

Trained with..	Tested with..	Angle	Angle+magn
gt t-flow	gt t-flow	77.47	93.23
gt t-flow	PWC t-flow	24.06	24.44
PWC t-flow	PWC t-flow	77.79	78.69

We trained four variants of our segmentation network with: (1) angle of the rotation-compensated flow of FT3D, (2) angle and magnitude of the rotation-compensated flow of FT3D (angle+magn), (3) angle of the estimated rotation-compensated flow, and (4) angle and magnitude of the estimated rotation-compensated flow. We show consistently better performance by including magnitude. The performance is the worst when the network is trained on the angle of the rotation-compensated ground truth flow. Here, the noise in angle leads to a very significant drop on estimated optical flow data. Segmentation accuracy is measured on the FT3D test set with intersection over union (IoU)

tion about independent object motion from the angle alone, and *does including the flow magnitude (training the network on the full optical flow) improve motion segmentation?*. We show this analysis in Table 5, with further variants of our network. Using angle and magnitude together (angle+magn in the table) leads to the best performance. However, note that we achieve reasonable segmentation quality even when using the flow angle alone. The network trained on ground truth optical flow adapts very poorly to estimated optical flow, with the segmentation accuracy dropping from 93.23% to 24.44% for the angle+magn variant.

Rotation estimation via likelihood maximization

We show results on the Sintel data set (Table 6), and compare our new likelihood optimization procedure with Bideau and Learned-Miller (2016b). The ground truth optical flow and focal length is provided, so an accurate estimate of the camera's rotation is possible. Our camera rotation estimation based on maximizing the flow likelihood shows consistently better results on the Sintel data set. More importantly, the performance gap gets significant when using estimated flow as input for camera motion estimation. Since our proposed optimization approach incorporates an explicit noise model together with a strategy for robust initialization of our likelihood optimization procedure, it is significantly more robust to noisy flow data. Figure 10 pictures the influence of the three starting points for our optimization procedure as described in Sect. 3.2. While the commonly best initialization for the camera motion estimate - the camera motion estimate of the previous frame - is best in most cases (see Fig. 10), it does not capture specifically difficult cases leading to an overall degraded performance. Starting only from a single initialization point drastically restricts the search the the best solution and even excludes options at greater distance, that still may be practically feasible. More distant solutions may arise due to sudden changes of the camera's motion direction or due to the ambiguity of camera rotation and translation in case of unknown depth. These two cases are precisely covered by the two additional camera motion initialization avoiding common challenges of local minima. We firstly consider motion direction in opposite direction to the past motion and incorporate information about the scene depth information in our joint estimation of camera motion and motion segmentation.

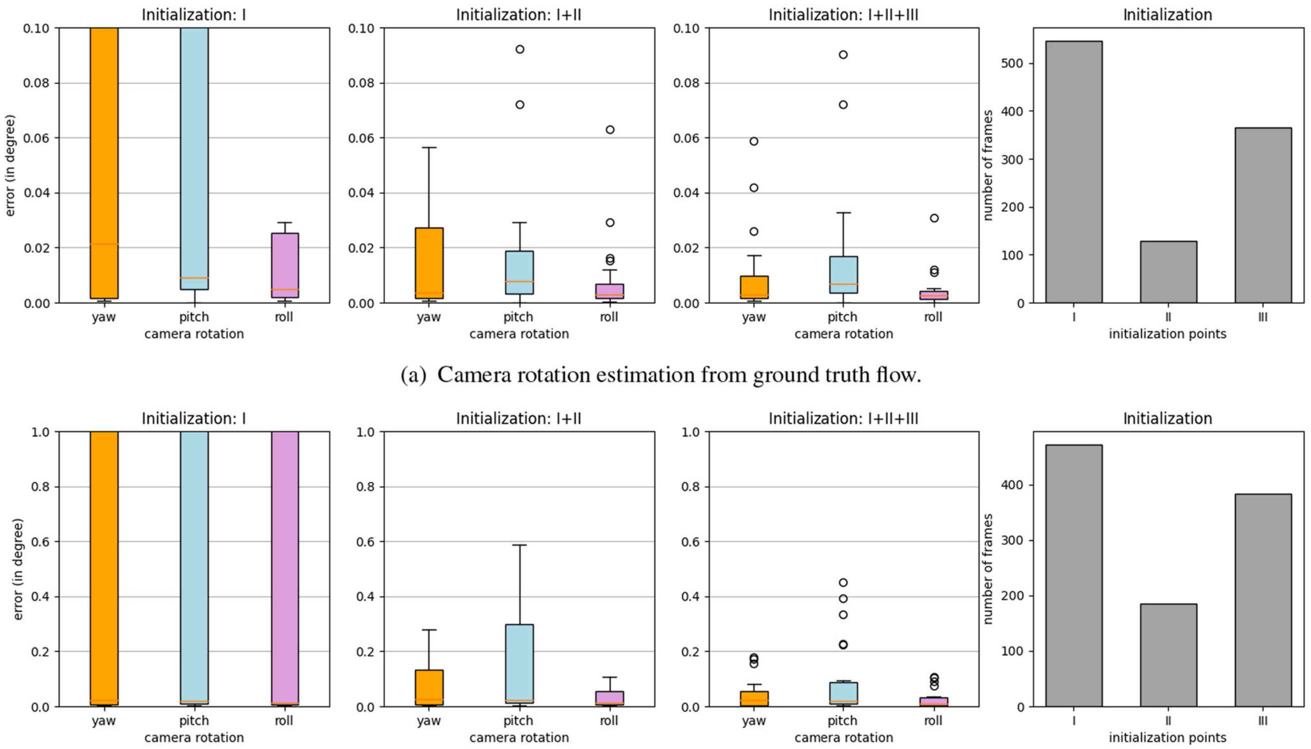
Table 6 Ablation study: Camera rotation estimation. Avg. yaw/pitch/roll error in degrees between two consecutive frames

	TMM	Ours
gt-flow	0.08 / 0.22 / 0.02	0.02 / 0.02 / 0.01
PWC-flow	0.13 / 0.34 / 0.04	0.04 / 0.09 / 0.02

gt-flow, PWC-flow: To evaluate rotation estimation we use ground-truth segmentation masks to weight the optim. loss. Thus, errors due to segmentation are not propagated throughout the video

5 Summary

In this work, we present a new approach that combines fundamental geometrical concepts and a learned understanding of moving objects. To this end new flow likelihood function is proposed to obtain a robust estimate of the camera motion. Knowing the camera motion removes complexity of estimated flow fields. It is shown, that rotation compensated flow fields containing only camera translation and object motion are more suitable to learn object motion patterns. This observation is supported by experiments on ideal flow data from a synthetic optical flow data set FlyingThings3D, as well as



(a) Camera rotation estimation from ground truth flow.

(b) Camera rotation estimation from estimated flow.

Fig. 10 Ablation study: Initialization of the optimization procedure over camera rotations. Avg. yaw/pitch/roll error in degrees between two consecutive frames. The influence of our three different starting points of our optimization procedure are evaluated: (I) camera rotation and translation estimate of previous frame, (II) camera rotation estimate weighted by estimated depth of the previous frame and translation estimate of previous frame and (III) camera rotation estimate weighted by estimated depth of the previous frame and the translation estimate of the previous frame in opposite direction. Please note the different scales of the error-axis showing the rotation error in degree. The bar plot pictures the discrete distribution over our three initialization points. Initialization (I) mostly leads to the minimum loss and thus is the choice in the majority of frames. *Left to right:* The average yaw/pitch/roll error in degree is shown for different initializations setups. We first initialize the optimization procedure with a single starting point - namely (I). We successively add starting points (II) and (III) and analyze the resulting error. While the average error decreases, also the amount of outliers is drastically reduced

on real world motion segmentation data sets such as DAVIS and MoCA.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-023-01859-x.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported in part by the ANR grant AVENUE (ANR-18-CE23-0011) and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2002/1 "Science of Intelligence" - project number 390523135.

Data Availability All data sets used in this article are publicly available. Data sharing is not applicable to this article as no new data sets were generated during the current study.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Bideau, P., & Learned-Miller, E. (2016a). A detailed rubric for motion segmentation. arXiv preprint arXiv:1610.10033 .

Bideau, P., & Learned-Miller, E. (2016b). It's moving! A probabilistic model for causal motion segmentation in moving camera videos. In *Proceedings of the European conference on computer vision*.

Bideau, P., Menon, R. R. & Learned-Miller, E. (2018). September. MoA-Net: Self-supervised motion segmentation. In *Proceedings of the European conference on computer vision workshops*.

Bideau, P., RoyChowdhury, A., Menon, R. R., & Learned-Miller, E. (2018). The best of both worlds: Combining CNNs and geometric constraints for hierarchical motion segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Brox, T., & Malik, J. (2010). Object segmentation by long term analysis of point trajectories. In *Proceedings of the European conference on computer vision*.

Butler, D. J., Wulff, J., Stanley, G. B., & Black, M. J. (2012). A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European conference on computer vision*.

Cheng, J., Tsai, Y. H., Wang, S., & Yang, M. H. (2017). SegFlow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*.

Choudhury, S., Karazija, L., Laina, I., Vedaldi, A., & Rupprecht, C. (2022). Guess what moves: Unsupervised video and image segmentation by anticipating motion. arXiv preprint arXiv:2205.07844 .

Dave, A., Tokmakov, P., & Ramanan, D. (2019). Towards segmenting anything that moves. In *Proceedings of the IEEE international conference on computer vision workshops*.

Ding, H., Liu, C., He, S., Jiang, X., Torr, P. H., & Bai, S. (2023). MOSE: A new dataset for video object segmentation in complex scenes. arXiv preprint arXiv:2302.01872 .

Eisner, B., Zhang, H., Held, D. (2022). FlowBot3D: Learning 3D articulation flow to manipulate articulated objects. In *Proceedings of robotics: Science and systems*, New York City, NY, USA.

Faktor, A., & Irani, M. (2014). Video segmentation by non-local consensus voting. In *Proceedings of the British machine vision conference*.

Fragkiadaki, K., Zhang, G., & Shi, J. (2012). Video segmentation by tracing discontinuities in a trajectory embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Gordon, A., Li, H., Jonschkowski, R., & Angelova, A. (2019). Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE international conference on computer vision*, pp. 8977–8986.

Horn, B. K. (1999). Projective geometry considered harmful. *Unpublished Memo* .

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Irani, M., & Anandan, P. (1998). A unified approach to moving object detection in 2D and 3D scenes. *IEEE Transactions Pattern Analysis and Machine Intelligence, 20*(6), 577–589.

Jain, S., Xiong, B., & Grauman. K. (2017). FusionSeg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Jin, Y., Tao, L., Di, H., Rao, N. I., & Xu, G. (2008). Background modeling from a free-moving camera by multi-layer homography algorithm. In Proceedings of the IEEE international conference on image processing

Ke, Q., & Kanade, T. (2002). *A robust subspace approach to layer extraction*. IEEE: In Workshop on motion and video computing.

Keuper, M. (2017). Higher-order minimum cost lifted multicuts for motion segmentation. In *Proceedings of the IEEE international conference on computer vision*.

Keuper, M., Andres, B., & Brox, T. (2015). Motion trajectory segmentation via minimum cost multicuts. In *Proceedings of the IEEE international conference on computer vision*.

Koh, Y. J., & Kim, C. S. (2017). Primary object segmentation in videos based on region augmentation and reduction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Lai, Z., Lu, E., & Xie, W. (2020). MAST: A memory-augmented self-supervised tracker. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Lamdouar, H., Xie, W., & Zisserman, A. (2021). Segmenting invisible moving objects. In *British machine vision conference*.

Lamdouar, H., Yang, C., Xie, W., & Zisserman, A. (2020). Betrayed by motion: Camouflaged object discovery via motion segmentation. In *Proceedings of the Asian conference on computer vision*.

Lezama, J., Alahari, K., Sivic, J., & Laptev, I. (2011). Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Longuet-Higgins, H.C., Prazdny, K., et al. (1980). The interpretation of a moving retinal image. *Proceedings of the royal society of London. Series B, Biological sciences 208*(1173): 385–397 .

Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., & Porikli, F. (2019). 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*.

Luiten, J., Voigtlaender, P., & Leibe, B. (2019). Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Proceedings of the Asian conference on computer vision 2018, Revised Selected Papers*, pp. 565–580. Springer.

Mahadevan, S., Athar, A., Ošep, A., Hennen, S., Leal-Taixé, L., & Leibe, B. (2020). Making a case for 3D convolutions for object segmentation in videos. In *Proceedings of the British machine vision conference*.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., & Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Narayana, M., Hanson, A., & Learned-Miller, E. (2013). Coherent motion segmentation in moving camera videos using optical flow orientations. In *Proceedings of the IEEE international conference on computer vision*.

Ochs, P., & Brox, T. (2011). Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions. In *Proceedings of the IEEE international conference on computer vision*. IEEE.

Ogale, A. S., Fermüller, C., & Aloimonos, Y. (2005). Motion segmentation using occlusions. *IEEE Transactions Pattern Analysis and Machine Intelligence, 27*(6), 988–992.

Papazoglou, A., & Ferrari, V. (2013). Fast object segmentation in unconstrained video. In *Proceedings of the IEEE international conference on computer vision*.

Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., & Sorkine-Hornung, A. (2016). A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Pirenne, M. H. (1952). The scientific basis of Leonardo da Vinci's theory of perspective. *The British Journal for the Philosophy of Science, 3*(10), 169–185.

Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., & Black, M. J. (2019). Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Shen, J., Peng, J., & Shao, L. (2018). Submodular trajectories for better motion segmentation in videos. *IEEE Transactions on Image Processing, 27*(6), 2688–2700.

Sun, D., Roth, S., & Black, M. J. (2010). Secrets of optical flow estimation and their principles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Sun, D., Yang, X., Liu, M. Y., & Kautz, J. (2018). PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Taylor, B., Karasev, V., & Soatto, S. (2015). Causal video object segmentation from persistence of occlusions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Tokmakov, P., Alahari, K., & Schmid, C. (2017a). Learning motion patterns in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Tokmakov, P., Alahari, K., & Schmid, C. (2017b). Learning video object segmentation with visual memory. In *Proceedings of the IEEE international conference on computer vision*.

Torr, P. H. S. (1998). Geometric motion segmentation and model selection. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 356*(1740), 1321–1340.

Tung, H. Y. F., Cheng, R., & Fragkiadaki, K. (2019). Learning spatial common sense with geometry-aware recurrent networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Vertens, J., Valada, A., & Burgard, W. (2017). SMSnet: Semantic motion segmentation using deep convolutional neural networks. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems*.

Vidal, R., & Ma, Y. (2004). A unified algebraic approach to 2-D and 3-D motion segmentation. In *Proceedings of the European conference on computer vision*.

Vidal, R., Soatto, S., Ma, Y., & Sastry, S. (2002). Segmentation of dynamic scenes from the multibody fundamental matrix. In *Proceedings of the European conference on computer vision workshops*.

Walls, G. (1962). The evolutionary history of eye movements. *Vision Research, 2*(1–4), 69–80.

Wang, J. Y., & Adelson, E. H. (1994). Representing moving images with layers. *IEEE Transactions on Image Processing, 3*(5), 625–638.

Wulff, J., Butler, D. J., Stanley, G. B., & Black, M. J. (2012). Lessons and insights from creating a synthetic optical flow benchmark. In *Proceedings of the European conference on computer vision workshops*.

Xiao, J., & Shah, M. (2005). Motion layer extraction in the presence of occlusion using graph cuts. *IEEE Transactions Pattern Analysis and Machine Intelligence, 27*(10), 1644–1659.

Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., & Huang, T. (2018). YouTube-VOS: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 .

Xu, X., Fah Cheong, L., & Li, Z. (2018). Motion segmentation by exploiting complementary geometric models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Xu, Z., Zhanpeng, H., & Song, S. (2022). UMPNet: Universal manipulation policy network for articulated objects. *IEEE robotics and automation letters*.

Yan, J., & Pollefeys, M. (2006). A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Proceedings of the European conference on computer vision*.

Yang, C., Lamdouar, H., Lu, E., Zisserman, A., & Xie, W. (2021). Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE international conference on computer vision*.

Yang, G., & Ramanan, D. (2021). Learning to segment rigid motions from two frames. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Yang, Y., Loquercio, A., Scaramuzza, D., & Soatto, S. (2019). Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Zamalieva, D., & Yilmaz, A. (2014). Background subtraction for the moving camera: A geometric approach. *Computer Vision and Image Understanding, 127*, 73–85.

Zhang, G., Jia, J., Xiong, W., Wong, T. T., Heng, P. A., & Bao, H. (2007). Moving object extraction with a hand-held camera. In *Proceedings of the IEEE international conference on computer vision*.

Zhou, T., Li, J., Wang, S., Tao, R., & Shen, J. (2020). MATNet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Transactions on Image Processing, 29*, 8326–8338.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.