

Multi-Object Discovery by Low-Dimensional Object Motion

Sadra Safadoust Fatma Güney

KUIS AI Center and Department of Computer Engineering, Koç University

ssafadoust20@ku.edu.tr fguney@ku.edu.tr

Abstract

Recent work in unsupervised multi-object segmentation shows impressive results by predicting motion from a single image despite the inherent ambiguity in predicting motion without the next image. On the other hand, the set of possible motions for an image can be constrained to a low-dimensional space by considering the scene structure and moving objects in it. We propose to model pixel-wise geometry and object motion to remove ambiguity in reconstructing flow from a single image. Specifically, we divide the image into coherently moving regions and use depth to construct flow bases that best explain the observed flow in each region. We achieve state-of-the-art results in unsupervised multi-object segmentation on synthetic and real-world datasets by modeling the scene structure and object motion. Our evaluation of the predicted depth maps shows reliable performance in monocular depth estimation.

1. Introduction

Finding objects on visual data is one of the oldest problems in computer vision, which has been shown to work to great extent in the presence of labeled data. Achieving it without supervision is important given the difficulty of obtaining pixel-precise masks for the variety of objects encountered in everyday life. In the absence of labels, motion provides important cues to group pixels corresponding to objects. The existing solutions use motion either as input to perform grouping or as output to reconstruct as a way of verifying the predicted grouping. The current methodology fails to incorporate the underlying 3D geometry creating the observed motion. In this work, we show that modeling geometry together with object motion significantly improves the segmentation of multiple objects without supervision.

Unsupervised multi-object discovery is significantly more challenging than the single-object case due to mutual occlusions. Therefore, earlier methods in unsupervised segmentation focused on separating a foreground object from the background whereas multi-object methods have been mostly limited to synthetic datasets or resorted to additional

supervision on real-world data such as sparse depth [15].

While sparse-depth supervision can be applied to driving scenarios [15], depth information is not typically available on common video datasets. Moreover, video segmentation datasets such as DAVIS [48, 49] contain a wide variety of categories under challenging conditions such as appearance changes due to lighting conditions or motion blur. The motion information can be obtained from video sequences via optical flow. Optical flow not only provides motion cues for grouping [64] but can also be used for training on synthetic data without suffering from the domain gap while transferring to real data [63]. The problems in optical flow prediction on real-world data can be mitigated to some extent by relating flow predictions from multiple frames [63].

In addition to problems in predicting optical flow, requiring flow as input prohibits the application of the method on static images. Another line of work [11, 33] uses motion for supervision at train time only. Based on the observation that objects create distinctive patterns in flow, initial work [11] fits a simple parametric model to the flow in each object region to capture the object motion. This way, the network can predict object regions that can potentially move coherently from a single image at test time. There is an inherent ambiguity in predicting motion from a single image. Therefore, the follow-up work [33] predicts a distribution of possible motion patterns to reduce this ambiguity. This also allows extending it to the multi-object case by mitigating the over-segmentation problem of the initial work [11].

In this work, we propose to model pixel-wise geometry to remove ambiguity in reconstructing flow from a single image. Optical flow is the difference between the 2D projections of the 3D world in consecutive time steps. By modeling the 3D geometry creating these projections, we directly address the mutual occlusion problem due to interactions of multiple objects. This problem has been crudely addressed by previous work with a depth-ordered layer representation [63]. Instead of assuming a single depth layer per object, we predict pixel-wise depth which provides more expressive power in explaining the observed motion. Furthermore, we do not use flow as input during inference, allowing us to evaluate our method on single-image datasets.

Recent work [5] showed that motion resides in a low-dimensional subspace, and its reconstruction can be used to supervise monocular depth prediction. Despite many possible flow fields, the space of possible flow fields is spanned by a small number of basis flow fields related to depth and independently moving objects. While [5] focuses on modeling camera motion for quantitatively evaluating depth in static scenes, it also points to the fact that the object motion can be similarly modeled in a low-dimensional subspace by simply masking the points in the object region. Given the difficulty of predicting pixel-wise masks, simple object embeddings are used to cluster independently moving objects. We instead predict the object regions jointly with depth to find the low-dimensional object motion that best explains the observed flow in each region.

Our approach works extremely well on synthetic Multi-Object Video (MOVi) datasets [23], significantly outperforming previous work, especially in more challenging MOVi-{C,D,E} partitions and performing comparably on visually simpler MOVi-A due to difficulty of estimating depth. We use motion only for supervision at train time, therefore our method can be successfully applied to still images of CLEVR [31] and CLEVRTEX [34] and shows state-of-the-art performance. More impressively, our method can segment multiple objects on real-world videos of DAVIS-2017 [49] from a single image at test time, exceeding the performance of the state-of-the-art that uses flow from multiple frames as input [63]. In addition to evaluating segmentation, we show that our method can also reliably predict depth in real-world self-driving scenarios on KITTI [21].

2. Related Work

Basis Learning. Early work showed that optical flow estimation due to camera motion can be constrained using a subspace formulation for flow [28]. Basis learning has been used as a regularization in low-level vision, unifying tasks such as depth, flow, and segmentation [55]. PCA-Flow [61] builds a higher dimensional flow subspace from movies to represent flow as a weighted sum of flow bases. Recent work [67] learns the coefficients to combine eight pre-defined flow bases for homography estimation.

Motion as Input. Most of the work in motion segmentation focuses on the single-object case. While earlier work uses traditional methods to cluster pixels into similar motion groups [6, 35, 47, 62], later methods train deep neural networks which take flow as input and predict segmentation as output [13, 57, 58]. Another work [66] uses the distinctiveness of motion in the case of foreground objects by proposing an adversarial setting to predict motion from context. Segmenting objects in camouflaged settings can be achieved by modeling background motion to remove its effect and highlight the moving foreground object [3, 4, 38].

Recent work uses consistency between two flow fields computed under different frame gaps for self-supervision [64].

The most relevant to our work is OCLR [63] which extends motion segmentation to multiple objects by relating motion extracted from multiple frames using a transformer in a layered representation. In this work, we show that better results can be achieved on real data even from a single image by modeling pixel-wise geometry.

Motion for Supervision. While using motion only as input works well where appearance fails, e.g. the camouflage datasets, RGB carries important information that might be missing in flow. DyStaB [65] trains a dynamic model by exploiting motion for temporal consistency and then uses it to bootstrap a static model which takes a single image as input. A single image network is used to predict a segmentation in [40] and then the motion of each segment is predicted with a two-frame motion network. While image warping loss is used in [40] for self-supervision, recent work [11, 33] uses flow reconstruction loss by assuming the availability of flow at train time only. GWM [11] segments foreground objects by fitting an approximate motion model to each segment and then merging them using spectral clustering. The follow-up work [33] extends it to multiple objects by predicting probable motion patterns for each segment with a distribution. We also reconstruct flow for supervision but differently, we account for 3D to remove the ambiguity in reconstructing motion from a single image.

The most relevant to our work is the previous work that uses flow as a source of supervision for depth [5] or segmentation [11, 34]. In this work, we model both depth and segmentation with supervision from motion.

Multi-Object Scene Decomposition. Our work is also related to scene decomposition approaches which are mostly evaluated on synthetic datasets. The earlier image-based decomposition approaches such as MONet [7] and IO-DINE [24] use a sequential VAE structure where the decomposition at a step can affect the remaining parts to be explained in the next step. GENESIS [17] follows an object-centric approach by accounting for component interactions, which is extended to more realistic scenarios with an autoregressive prior in the follow-up work [18]. Slot Attention [42] uses an iterative attention mechanism to decompose the image into a set of slot representations. A hierarchical VAE is used in [16] to extract symmetric and disentangled representations.

There are also video-based approaches to multi-object scene decomposition. SCALOR [30] focuses on scaling generative approaches to crowded scenes in terms of object density. SIMONE [32] learns a factorized latent space to separate object semantics that is constant in the sequence from the background which changes at each frame according to camera motion. SAVi [36] extends Slot Attention [42]

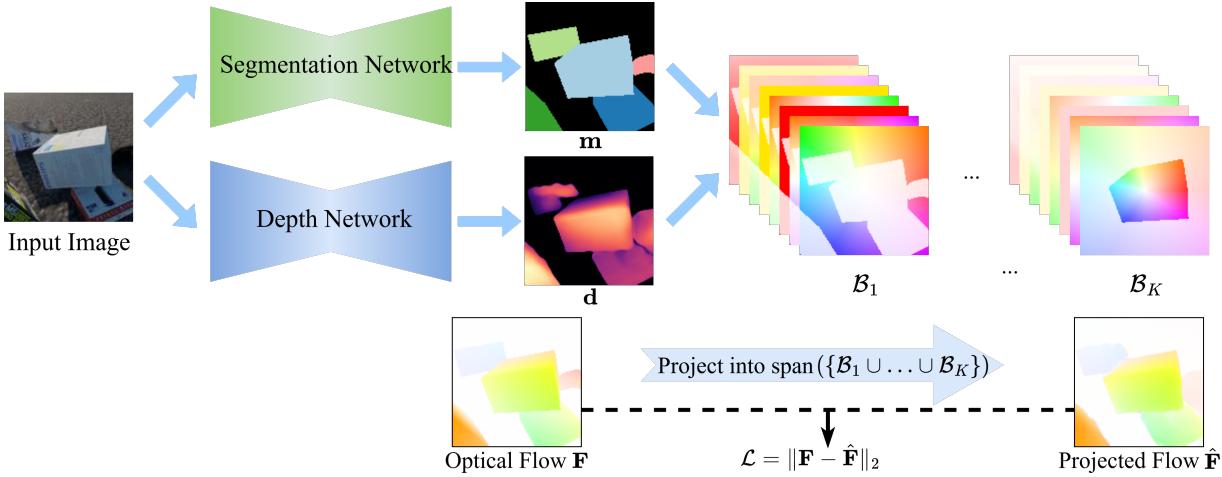


Figure 1: Overview of our Approach. From a single image, we use a segmentation and a depth network to predict a segmentation mask m and a disparity map d . Based on these predictions, we construct the bases for the space of the possible optical flows for K distinctly moving regions on the image. Each moving region i is represented with a separate basis \mathcal{B}_i . Given optical flow \mathbf{F} as input, either ground truth or estimated by an off-the-shelf method, we project it into $\text{span}(\bigcup_{i=1}^K \mathcal{B}_i)$. We use the distance between the input flow \mathbf{F} and the projected flow $\hat{\mathbf{F}}$ to supervise depth and segmentation. During inference, our networks can be used to predict depth and segmentation from a single image.

to videos and SAVi++ [15] extends it to real-world driving scenarios with sparse depth supervision.

Self-Supervised Monocular Depth Estimation. Zhou et al. [69] train a pose network to estimate the pose between the frames in a sequence and jointly train it with the depth network. Godard et al. [22] improves the results with a better loss function and other design choices. Guizilini et al. [25] learn detail-preserving representations using 3D packing and unpacking blocks. Given instance segmentation masks, a line of work [9, 39] models the motion of objects in the scene in addition to the camera motion to go beyond the static-scene assumption. While the object masks are supervised using ground truth masks in [37], the masks are learned without supervision as an auxiliary output in [53] for better depth estimation. While they require multiple frames during inference, our approach can estimate masks from a single image. Additionally, our method does not use camera intrinsics.

3. Depth-Aware Multi-Object Segmentation

The observed motion in 2D is the result of 3D scene structure and independently moving objects. By predicting the scene structure in terms of depth and locating independently moving objects, we can accurately reconstruct the optical flow corresponding to the observed motion in 2D. Towards this purpose, we use a low-dimensional parameterization of optical flow based on depth (Section 3.1). In this low-dimensional representation, we can accurately reconstruct flow from a rigid motion. We extend this parametriza-

tion to a number of rigidly moving objects to find the regions corresponding to objects (Section 3.2). See Fig. 1 for an overview of our approach.

3.1. Low-Dimensional Motion Representation

The space of all possible optical flow fields is very high-dimensional, i.e. in $\mathbb{R}^{H \times W \times 2}$. However, conditioned on the scene structure, only a small fraction of all flow fields are possible. Previous work [27, 5] has shown that the set of possible instantaneous flows for a moving camera in a static scene forms a linear space with six basis vectors:

$$\mathcal{B}_0 = \{\mathbf{b}_{\mathbf{T}x}, \mathbf{b}_{\mathbf{T}y}, \mathbf{b}_{\mathbf{T}z}, \mathbf{b}_{\mathbf{R}x}, \mathbf{b}_{\mathbf{R}y}, \mathbf{b}_{\mathbf{R}z}\} \quad (1)$$

These basis vectors correspond to translation and rotation along the x , y , and z axes, respectively. For an image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the values of each basis vector $\mathbf{b}_i \in \mathbb{R}^{H \times W \times 2}$ for a given pixel (u, v) can be calculated as follows:

$$\begin{aligned} \mathbf{b}_{\mathbf{T}x} &= \begin{bmatrix} f_x & d \\ 0 & 0 \end{bmatrix}, & \mathbf{b}_{\mathbf{R}x} &= \begin{bmatrix} f_y^{-1} & \bar{u} & \bar{v} \\ f_y + f_y^{-1} & \bar{v}^2 & 0 \end{bmatrix} \\ \mathbf{b}_{\mathbf{T}y} &= \begin{bmatrix} 0 & d \\ f_y & d \end{bmatrix}, & \mathbf{b}_{\mathbf{R}y} &= \begin{bmatrix} f_x + f_x^{-1} & \bar{u}^2 \\ f_x^{-1} & \bar{u} & \bar{v} \end{bmatrix} \\ \mathbf{b}_{\mathbf{T}z} &= \begin{bmatrix} -\bar{u} & d \\ -\bar{v} & d \end{bmatrix}, & \mathbf{b}_{\mathbf{R}z} &= \begin{bmatrix} f_x & f_y^{-1} & \bar{v} \\ -f_y & f_x^{-1} & \bar{u} \end{bmatrix} \end{aligned} \quad (2)$$

where f_x, f_y are the focal lengths of the camera. For brevity, we define $\bar{u} = u - c_x$ and $\bar{v} = v - c_y$ to be the centered pixel coordinates according to (c_x, c_y) , the principal point of the camera. With a slight abuse of notation, we

do not write the basis vectors as a function of (u, v) and use d to denote the disparity $\mathbf{d}(u, v)$ at a pixel (u, v) .

We train a monocular depth network to predict inverse depth, disparity \mathbf{d} from a single image. Then, the predicted disparity at each pixel is used to form the translation part of the basis vectors as shown in Eq. (2). Note that predicted disparity values do not affect the rotation but form the low-dimensional motion representation via translation. The depth network receives gradients directly from the flow reconstruction loss as explained next in Section 3.2.

In Eq. (2), camera parameters including the principal point (c_x, c_y) and the focal lengths f_x, f_y are required to calculate the basis vectors. We assume the principal points to be at the center of the image. However, we do not assume to know the values of focal lengths. Instead, we only assume that $f_x = f_y$. In this case, as demonstrated by [5], we can rewrite \mathcal{B}_0 as a set of 8 vectors that do not depend on the values of f_x and f_y . For more details, please see Supplementary. Even without knowing the actual values of the focal lengths, we can obtain quite accurate depth predictions with supervision from flow (Section 4).

3.2. Segmentation by Object Motion

We extend the formulation introduced in Section 3.1 to handle the instantaneous flow from multiple object motion. As stated in [5], for a rigidly moving object in the scene, there is an equivalent camera motion. Therefore the space of optical flow from a rigidly moving object is the same as the space of optical flow from camera motion restricted to points in the object. Consider a scene with K regions corresponding to moving parts including the background and multiple objects. If we represent each region $i \in \{1, \dots, K\}$ with ones on a binary mask $\mathbf{m}_i \in \{0, 1\}^{H \times W \times 1}$, then a basis for the space of possible flows can be defined as follows:

$$\mathcal{B} = \{\mathcal{B}_1 \cup \mathcal{B}_2 \cup \dots \cup \mathcal{B}_K\} \quad (3)$$

where \mathcal{B}_i refers to the basis for the space of possible flows restricted to region i as:

$$\mathcal{B}_i = \{\mathbf{m}_i \mathbf{b} \mid \mathbf{b} \in \mathcal{B}_0\}. \quad (4)$$

We train a segmentation network to divide the image into coherently moving regions $\mathbf{m} \in [0, 1]^{H \times W \times K}$, representing soft assignments over K regions. We use the predicted mask $\mathbf{m}_i \in [0, 1]^{H \times W \times 1}$ of the region i to obtain the basis corresponding to that region according to Eq. (4).

Training Objective. Based on the predicted disparity map \mathbf{d} and the segmentation map \mathbf{m} , we form the basis \mathcal{B} for the space of possible optical flows for the image according to Eq. (3) and Eq. (4). We denote the optical flow where the input image is the source frame as $\mathbf{F} \in \mathbb{R}^{H \times W \times 2}$. It can be either ground truth flow or the output of a two-frame flow

network such as RAFT [56]. We project \mathbf{F} into the space spanned by \mathcal{B} in a differentiable manner to obtain $\hat{\mathbf{F}}$. For the details of the projection, please refer to Supplementary. We define our loss function as the L_2 distance between the given flow \mathbf{F} and the reconstructed flow $\hat{\mathbf{F}}$ and use it to train depth and segmentation networks jointly:

$$\mathcal{L} = \|\mathbf{F} - \hat{\mathbf{F}}\|_2 \quad (5)$$

4. Experiments

4.1. Datasets

Synthetic Datasets. For comparison to image-based methods, we evaluate our method on the CLEVR [31] and CLEVRTEX [34] datasets. CLEVR is a dataset of still images depicting multiple objects of random shape, size, color, and position. CLEVRTEX is similar to CLEVR but contains more diverse textures and shapes. Because our method needs optical flow for training, we train our model on the MOVINGCLEVR and MOVINGCLEVRTEX datasets [33], which are video extensions of CLEVR and CLEVRTEX scenes. We train on the video versions but evaluate on the original test sets of CLEVR and CLEVRTEX.

For comparison to video-based methods, we use the Multi-Object Video (MOV) datasets [23]. Similar to [33], we use the MOV-{A, C, D, E} variants. MOV-A is based on CLEVR [31] and contains scenes with a static camera and multiple objects with simple textures and uniform colors tossed on a gray floor. MOV-C is more challenging due to realistic everyday objects with rich textures on a more complex background. MOV-D increases the complexity by increasing the number of objects. MOV-E is even more challenging as it features camera motion as well.

In all our experiments on synthetic datasets, we use a resolution of 128×128 and the ground truth optical flow.

Real-World Datasets. We use the common video segmentation dataset DAVIS-2017 [49] containing 90 video sequences where each sequence has one or more moving objects. We follow the evaluation protocol of [63] where the ground truth objects are reannotated by assigning the same label to the jointly moving objects. We resize the images to a resolution of 128×224 during training and use the flow from RAFT [56] with $\{-8, -4, 4, 8\}$ gaps between frames.

Additionally, we evaluate our method on the KITTI driving dataset [21, 20]. Following [2], we train on the whole training set and evaluate the segmentation results on the instance segmentation benchmark consisting of 200 frames. We use a resolution of 128×416 and the flow from RAFT [56] with a gap of +1. Additionally, we evaluate our depth results on KITTI. Following prior work [69, 22], we evaluate depth on the Eigen split [14] of the KITTI dataset using improved ground truth [59] to be comparable to self-supervised monocular depth estimation approaches.

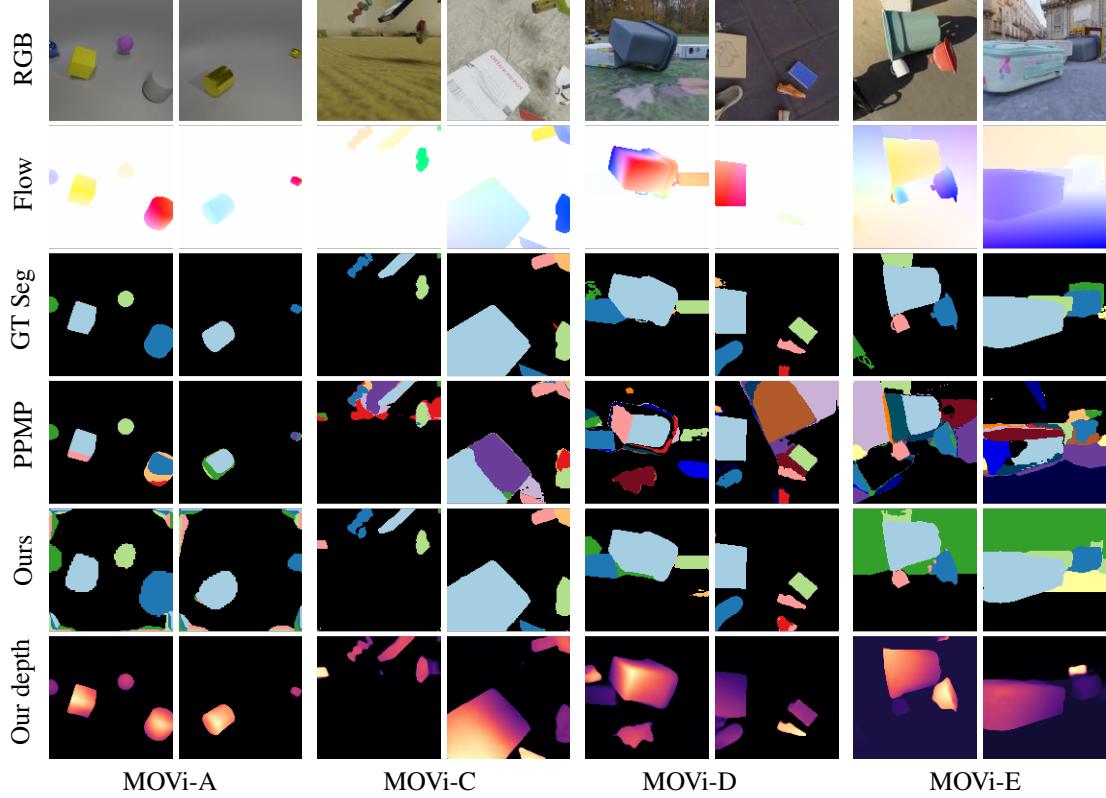


Figure 2: **Visualization of Depth and Segmentation Results on MOVi datasets.** Our method performs accurate segmentations, while PPMP suffers from over-segmentation and also mistakenly segments parts of the background as objects .

4.2. Architecture Details

We use the same architecture used in [50] for depth and Mask2Former [10] for segmentation, using only the segmentation head. We use different backbones for the segmentation network on the synthetic and real datasets. On synthetic datasets, we follow [33, 36, 42] and utilize a 6-layer CNN. On real-world datasets, following [11], we use a ViT-B/8 transformer pre-trained self-supervised using DINO [8] on ImageNet [52]. On all of the datasets, we use 6 object queries in the segmentation network, which translates to $K = 6$, except for CLEVR, where we use $K = 8$.

We use a fixed learning rate of 5×10^{-5} for the depth network and use 1.5×10^{-4} with a linear warm-up for the first 5K iterations for the segmentation network, reduced to 1.5×10^{-5} after 200K iterations. We train both networks with AdamW optimizer [43] for 250K iterations. See Supplementary for further details, we will also share the code.

4.3. Evaluation Details

Metrics. Following prior work [34, 36, 33], we evaluate segmentation on synthetic datasets using Adjusted Rand Index on foreground pixels (FG-ARI) and mean Intersection over Union (mIoU). ARI measures how well predicted and

ground truth segmentation masks match in a permutation-invariant manner. For mIoU, we first apply Hungarian matching and calculate the mean over the maximum between the number of ground-truth and predicted segments.

On DAVIS-2017 [49], we use the standard \mathcal{J} , \mathcal{F} metrics and perform the Hungarian matching per frame, similar to other datasets. Note that we focus on the multi-object segmentation task without using any labels for segmentation at train or test time. For KITTI, we use the FG-ARI metric, following [33, 2]. For evaluating depth, we use the standard metrics used in monocular depth estimation [14, 19].

Post-processing. We also report the results on segmentation using the post-processing method introduced in [33]. They extract the connected components in the model output and choose the largest K masks and discard any masks that take up less than 0.1% of the image. Then they combine the discarded masks with the largest mask. The results with post-processing are marked with \dagger in the tables.

4.4. Results on Synthetic Datasets

We evaluate our method on synthetic datasets and compare its performance to both image-based and video-based methods. Our method uses motion during training only. Therefore, it can also be evaluated on the image datasets.

	MOVi-A		MOVi-C		MOVi-D		MOVi-E	
	FG-ARI↑	mIoU↑	FG-ARI↑	mIoU↑	FG-ARI↑	mIoU↑	FG-ARI↑	mIoU↑
GWM [10]	70.30	42.27	49.98	30.17	39.78	18.38	42.50	18.74
SCALOR [30]	59.57	44.41	40.43	22.54	-	-	-	-
SAVi [36]	88.30	62.69	43.26	31.92	43.45	10.60	17.39	5.75
PPMP [33]	84.01	60.08	61.18	34.72	55.74	23.50	62.62	25.78
PPMP [†]	85.41	<u>76.19</u>	61.24	37.26	55.18	25.21	63.11	28.59
PPMP [†] (Swin)	90.08	84.76	67.67	52.17	66.41	30.40	72.73	35.30
Ours	56.09	36.48	<u>73.80</u>	<u>54.48</u>	<u>76.41</u>	<u>58.82</u>	<u>78.33</u>	<u>47.38</u>
Ours [†]	70.15	46.26	74.64	59.24	77.15	59.68	80.83	50.48

Table 1: **Segmentation Results on MOVi Datasets.** The best result in each column is shown in **bold**, and the second best is underlined. [†] indicates post-processing, and (Swin) denotes using a Swin transformer as the backbone.

Video-Based Methods. We compare our method to other video-based methods on the MOVi video datasets in Table 1 and Fig. 2. All of the methods in Table 1 use optical flow for supervision. Differently, SCALOR [30] and SAVi [36] use all frames in a video, whereas PPMP [33] and our method perform single-image segmentation, one frame at a time without using any motion information at test time. On the simpler MOVi-A dataset, the performance of our method falls behind SAVi [36] and PPMP [33]. PPMP with Swin transformer [41] performs the best overall. With the same 6-layer CNN backbone and without post-processing, SAVi performs the best. Despite the advantage of motion information, the success of SCALOR [30] and SAVi [36] is limited to visually simpler MOVi-A.

On the more challenging MOVi-{C,D,E} datasets, our method, even without post-processing, significantly outperforms all the previous methods in both metrics, with or without post-processing. The previous state-of-the-art, PPMP [33] uses the same backbone in their segmentation network as ours. Even with a more powerful backbone (Swin transformer [41]) and post-processing, the results of PPMP are still far behind our results without any post-processing. From MOVi-C to MOVi-E, the performance gap between our method and the others increases as the complexity of the dataset increases. Please see Supplementary for qualitative results with post-processing and evaluation of our estimated depth for objects in these datasets.

Image-Based Methods. We compare our method to other image-based methods on the CLEVR and CLEVRTEX datasets in Table 2. Our method outperforms the state-of-the-art method PPMP [33] in all metrics on both datasets except for mIoU on CLEVR without postprocessing and mIoU on CLEVRTEX with postprocessing. We point to **+9.01** improvement in mIoU on the more challenging CLEVRTEX dataset without postprocessing. See Supplementary for qualitative results on CLEVR and CLEVRTEX.

4.5. Results on Real-World Datasets

We compare our method to multi-object segmentation methods on real-world datasets including driving scenarios on KITTI and unconstrained videos on DAVIS-2017.

Results on DAVIS. Our method is the first image-based method to report performance in multi-object segmentation without using any labels during training or testing on

	CLEVR		CLEVRTEX	
	FG-ARI↑	mIoU↑	FG-ARI↑	mIoU↑
SPAIR [12]	77.13	<u>65.95</u>	0.00	0.00
MN [54]	72.12	56.81	38.31	10.46
MONet [7]	54.47	30.66	36.66	19.78
SA [42]	95.89	36.61	62.40	22.58
IODINE [24]	93.81	45.14	59.52	29.17
DTI-S [46]	59.54	48.74	79.90	33.79
GNM [29]	65.05	59.92	53.37	42.25
SAVi [36]	-	-	49.54	31.88
PPMP [32]	91.69	66.70	90.80	<u>55.07</u>
Ours	95.03	63.36	94.66	64.08
SPAIR [†] [12]	77.05	66.87	0.00	0.00
MN [†] [54]	72.08	57.61	38.34	10.34
MONet [†] [7]	61.36	45.61	35.64	23.59
SA [†] [42]	94.88	37.68	61.60	21.96
IODINE [†] [24]	93.68	44.20	60.63	29.40
DTI-S [†] [46]	89.86	53.38	79.86	32.20
GNM [†] [29]	65.67	63.38	53.38	44.30
PPMP [†] [33]	<u>95.94</u>	<u>84.86</u>	<u>92.61</u>	77.67
Ours [†]	96.95	86.38	95.32	<u>70.28</u>

Table 2: **Segmentation Results on CLEVR and CLEVRTEX Datasets.** The lower part with [†] shows the results with post-processing. The best result in each column is shown in **bold**, and the second best is underlined.

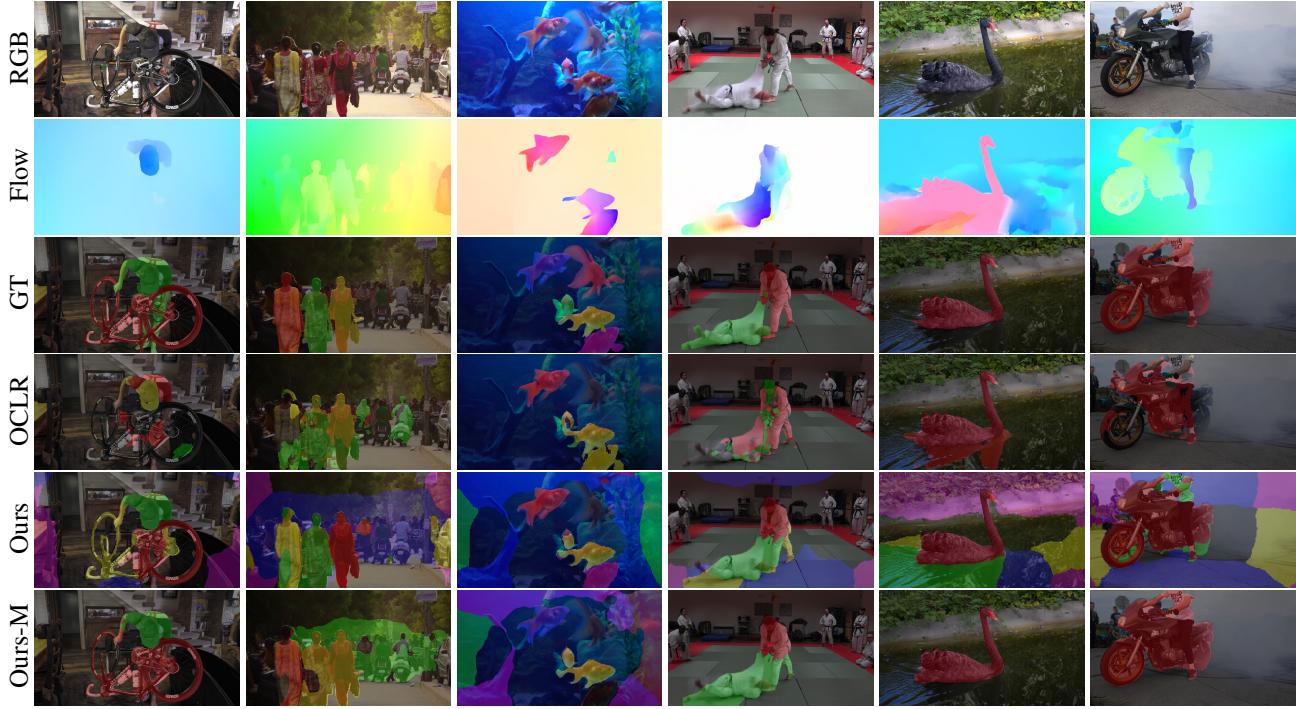


Figure 3: **Qualitative Comparison on DAVIS-2017.** While OCLR [63] misses some objects completely and suffers from relying on only optical flow as input, our method can segment a wide variety of multiple objects in everyday scenes.

DAVIS-2017. Therefore, in Table 3, we compare it to video-based approaches which use motion as input. We also compare to a simple baseline proposed in [63] based on Mask R-CNN [26] using optical flow as input. We use the labels re-annotated by [63] for evaluation, as explained in Section 4.1. Motion Grouping refers to [64] trained on DAVIS-2017. Motion Grouping (sup.), Mask R-CNN (flow) and OCLR are models trained on synthetic data from [63] in a supervised way using optical flow as input. Our method, which uses a single RGB image as input at test time, outperforms previous methods, including the state-of-the-art OCLR [63] that uses flow from multiple time steps. Ours-M refers to a version of our model where we adapt the spectral clustering approach of [11] to merge our predicted regions into the ground truth number of regions in each frame.

Model	$\mathcal{J} \& \mathcal{F} \uparrow$	$\mathcal{J} \uparrow$	$\mathcal{F} \uparrow$
Motion Grouping [64]	35.8	38.4	33.2
Motion Grouping (sup.)	39.5	44.9	34.2
Mask R-CNN (flow)	50.3	50.4	50.2
OCLR [63]	55.1	54.5	55.7
Ours	55.3	55.3	55.3
Ours-M	59.2	59.3	59.2

Table 3: **Multi-Object Segmentation Results on DAVIS-2017.** We evaluate by using the motion labels from [63].

Although not necessary to achieve state-of-the-art, this improves the results significantly.

We visualize the results of our model in comparison to OCLR [63] in Fig. 3. Our method can correctly segment a wide variety of objects such as the bike and the person in the first column and the multiple fish in the third, multiple people walking or fighting in the second and fourth. OCLR is highly affected by the inaccuracies in flow, unlike our method, as can be seen from the last two columns.

Results on KITTI. Since KITTI has ground truth depth, we evaluate our method in terms of both segmentation and monocular depth prediction on KITTI. The segmentation results on KITTI are presented in Table 4a. Our method is among the top-performing methods, outperforming earlier approaches. In addition to segmentation, our method can also predict depth from a single image. We present the evaluation of our depth predictions in comparison to self-supervised monocular depth estimation methods in Table 4b. Our depth network can predict reliable depth maps even without camera intrinsics with comparable results to recent self-supervised monocular depth estimation approaches [22, 25] that are specifically designed for that task and that use camera intrinsics.

We visualize our segmentation and depth results on KITTI in Supplementary. Our method can segment multiple moving objects such as car, bike, bus without using any semantic labels for training or motion information at test

Model	FG-ARI ↑	Abs Rel ↓	Sq Rel ↓	RSME ↓	RMSE log ↓	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
SA [42]	13.8	Zhou et al. [69]	0.176	1.532	6.129	0.244	0.758	0.921
MONet [7]	14.9	Mahjourian et al. [45]	0.134	0.983	5.501	0.203	0.827	0.944
SCALOR [30]	21.1	GeoNet [68]	0.132	0.994	5.240	0.193	0.833	0.953
S-IODINE [24]	14.4	DDVO [60]	0.126	0.866	4.932	0.185	0.851	0.958
MCG [1]	40.9	Ranjan et al. [51]	0.123	0.881	4.834	0.181	0.860	0.959
Ours	42.3	EPC++ [44]	0.120	0.789	4.755	0.177	0.856	0.961
Bao et al. [2]	47.1	Ours	0.107	1.539	4.027	0.149	0.911	0.971
PPMP [33]	50.8	Monodepth2 [22]	0.090	0.545	3.942	0.137	0.914	0.983
		PackNet-SfM [25]	0.078	0.420	3.485	0.121	0.934	0.996

(a) Segmentation

(b) Depth

Table 4: **Results on KITTI.** We evaluate the segmentation on the instance segmentation benchmark, and the depth on the KITTI Eigen split [14] with improved ground truth [59].

	MOVi-A		MOVi-C		MOVi-D		MOVi-E	
	FG-ARI↑	mIoU↑	FG-ARI↑	mIoU↑	FG-ARI↑	mIoU↑	FG-ARI↑	mIoU↑
Only-T	31.13	17.72	69.69	25.84	68.68	22.74	64.69	28.26
Only-R	90.43	75.63	68.51	53.00	70.57	56.97	73.03	40.34
Full	56.09	36.48	73.80	54.48	76.41	58.82	78.33	47.38
Only-T [†]	47.96	29.92	69.94	26.31	68.66	22.76	69.94	30.71
Only-R [†]	92.09	84.60	69.23	55.54	71.20	58.41	76.20	42.96
Full [†]	70.15	46.26	74.64	59.24	77.15	59.68	80.83	50.48

Table 5: **Ablation Study.** We perform an ablation study by using only the translation (Only-T) or rotation (Only-R) component and compare it to our model with both (Full). See text for details.

time. Furthermore, it can predict high-quality depth, capturing thin structures and sharp boundaries around objects.

4.6. Ablation Study

To evaluate the contribution of different types of flow basis, we perform an experiment by considering only translation or rotation and compare it to the full model on MOVi datasets in Table 5. Note that in the rotation-only case (Only-R), the depth predictions are not used and the depth network is not trained. Overall, the rotation-only model outperforms the translation-only model and the full model with both rotation and translation works the best on MOVi-{C, D, E} datasets with reliable depth predictions.

The trend is different on the simpler MOVi-A dataset. Only-R outperforms all models including the state-of-the-art in Table 1. We found that in the translation-only case, the depth cannot be predicted on MOVi-A due to missing texture and enough detail for the depth network to learn a mapping from a single image to depth. The rotation-only model, on the other hand, learns to group pixels in a region, based on their rotational motion which does not depend on depth. This ability explains the success of Only-R on simpler datasets. The importance of pixel-wise depth increases with the complexity of the dataset. On MOVi-E, for example, which has the most complex setup with a large number of objects and camera motion, predicting depth, from Only-R to Full, improves the performance the most.

5. Conclusion and Future Work

We presented a motion-supervised approach for multi-object segmentation that can work with a single RGB image at test time, therefore still applicable to image datasets. Our method is the first to consider geometry to remove ambiguity for multi-object segmentation from a single image without using any labels for segmentation. Modeling geometry significantly advances the state-of-the-art on commonly used synthetic datasets. We also evaluated our method on real-world datasets. Our method is the first image-based multi-object segmentation method to report state-of-the-art results on DAVIS-2017 without using motion at test time. We also report comparable results for depth prediction on KITTI and MOVi datasets where depth can be evaluated.

Predicting objects that can potentially move independently from a single image requires observing examples of various objects moving in the training set. Moreover, static objects send a mixed signal to the model. The coherent changes in the flow can be captured with the help of geometry as shown in our work. The remaining uncertainty can be addressed with a probabilistic formulation as done in previous state-of-the-art [33]. Another problem is scenes without enough information to predict depth as we observed on textureless MOVi-A. However, the lack of information to this extent rarely happens on real-world data.

Acknowledgements. Sadra Safadoust was supported by KUIS AI Fellowship and UNVEST R&D Center.

References

- [1] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014. 8
- [2] Zhipeng Bao, Pavel Tokmakov, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Discovering objects that can move. In *CVPR*, pages 11789–11798, 2022. 4, 5, 8
- [3] Pia Bideau and Erik G. Learned-Miller. It’s moving! a probabilistic model for causal motion segmentation in moving camera videos. In *ECCV*, 2016. 2
- [4] Pia Bideau, Aruni RoyChowdhury, Rakesh R. Menon, and Erik Learned-Miller. The best of both worlds: Combining cnns and geometric constraints for hierarchical motion segmentation. In *CVPR*, 2018. 2
- [5] Richard Strong Bowen, Richard Tucker, Ramin Zabih, and Noah Snavely. Dimensions of motion: Monocular prediction through flow subspaces. In *3DV*, 2022. 2, 3, 4
- [6] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 2
- [7] Christopher P. Burgess, Loïc Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matthew M. Botvinick, and Alexander Lerchner. MONet: Unsupervised scene decomposition and representation. *arXiv.org*, 1901.11390, 2019. 2, 6, 8
- [8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 5
- [9] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *AAAI*, 2019. 3
- [10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 5, 6
- [11] Subhabrata Choudhury, Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Guess What Moves: Unsupervised Video and Image Segmentation by Anticipating Motion. In *BMVC*, 2022. 1, 2, 5, 7
- [12] Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. In *AAAI*, 2019. 6
- [13] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting anything that moves. In *ICCV Workshops*, 2019. 2
- [14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 4, 5, 8
- [15] Gamaleldin F. Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C. Mozer, and Thomas Kipf. SAVI++: Towards end-to-end object-centric learning from real-world videos. In *NeurIPS*, 2022. 1, 3
- [16] Patrick Emami, Pan He, Sanjay Ranka, and Anand Rangarajan. Efficient iterative amortized inference for learning symmetric and disentangled multi-object representations. In *ICML*, 2021. 2
- [17] Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: generative scene inference and sampling with object-centric latent representations. In *ICLR*, 2020. 2
- [18] Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. GENESIS-V2: inferring unordered object representations without iterative refinement. In *NeurIPS*, 2021. 2
- [19] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 5
- [20] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *IJRR*, 2013. 4
- [21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2, 4
- [22] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 3, 4, 7, 8
- [23] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. In *CVPR*, 2022. 2, 4
- [24] Klaus Greff, Raphael Lopez Kaufman, Rishabh Kabra, Nicholas Watters, Christopher P. Burgess, Daniel Zoran, Loïc Matthey, Matthew M. Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *ICML*, 2019. 2, 6, 8
- [25] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 3, 7, 8
- [26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. In *ICCV*, 2017. 7
- [27] David J. Heeger and Allan D. Jepson. Subspace methods for recovering rigid motion I: Algorithm and implementation. *IJCV*, 1992. 3
- [28] M. Irani. Multi-frame optical flow estimation using subspace constraints. In *ICCV*, 1999. 2
- [29] Jindong Jiang and Sungjin Ahn. Generative neurosymbolic machines. In *NeurIPS*, 2020. 6
- [30] Jindong Jiang, Sepehr Janghorbani, Gerard de Melo, and Sungjin Ahn. SCALOR: generative world models with scalable object representations. In *ICLR*, 2020. 2, 6, 8
- [31] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 2, 4
- [32] Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loïc Matthey, Antonia Creswell, Matthew M. Botvinick, Alexander Lerchner, and Christopher P. Burgess. SIMONE: view-invariant,

- temporally-abstracted object representations via unsupervised video decomposition. In *NeurIPS*, 2021. 2, 6
- [33] Laurynas Karazija, Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised Multi-object Segmentation by Predicting Probable Motion Patterns. In *NeurIPS*, 2022. 1, 2, 4, 5, 6, 8
- [34] Laurynas Karazija, Iro Laina, and C. Rupprecht. ClevrTex: A texture-rich benchmark for unsupervised multi-object segmentation. In *NeurIPS*, 2021. 2, 4, 5
- [35] M. Keuper, B. Andres, and T. Brox. Motion trajectory segmentation via minimum cost multicut. In *ICCV*, 2015. 2
- [36] Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional Object-Centric Learning from Video. In *ICLR*, 2022. 2, 5, 6
- [37] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *ECCV*, 2020. 3
- [38] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *ACCV*, 2020. 2
- [39] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon. Learning monocular depth in dynamic scenes via instance-aware projection consistency. In *AAAI*, 2021. 3
- [40] Runtao Liu, Zhirong Wu, Stella X. Yu, and Stephen Lin. The emergence of objectness: Learning zero-shot segmentation from videos. In *NeurIPS*, 2021. 2
- [41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 6
- [42] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, 2020. 2, 5, 6, 8
- [43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv.org*, 1711.05101, 2017. 5
- [44] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *PAMI*, 42(10):2624–2641, 2019. 8
- [45] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2018. 8
- [46] Tom Monnier, Elliot Vincent, Jean Ponce, and Mathieu Aubry. Unsupervised Layered Image Decomposition into Object Prototypes. In *ICCV*, 2021. 6
- [47] Peter Ochs and Thomas Brox. Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions. In *ICCV*, 2011. 2
- [48] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 1
- [49] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbelaez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1, 2, 4, 5
- [50] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *PAMI*, 44(3):1623–1637, 2020. 5
- [51] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, 2019. 8
- [52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015. 5
- [53] Sadra Safadoust and Fatma Güney. Self-supervised monocular scene decomposition and depth estimation. In *3DV*, 2021. 3
- [54] Dmitriy Smirnov, Michael Gharbi, Matthew Fisher, Vitor Guizilini, Alexei A. Efros, and Justin Solomon. MarioNette: Self-supervised sprite learning. In *NeurIPS*, 2021. 6
- [55] Chengzhou Tang, Lu Yuan, and Ping Tan. Lsm: Learning subspace minimization for low-level vision. In *CVPR*, 2020. 2
- [56] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 4
- [57] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *CVPR*, 2017. 2
- [58] Pavel Tokmakov, Cordelia Schmid, and Alahari Karteek. Learning to segment moving objects. *IJCV*, 2019. 2
- [59] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant CNNs. In *THREEDV*, 2017. 4, 8
- [60] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018. 8
- [61] Jonas Wulff and Michael J. Black. Efficient sparse-to-dense optical flow estimation using a learned basis and layers. In *CVPR*, 2015. 2
- [62] Christopher Xie, Yu Xiang, Dieter Fox, and Zaid Harchaoui. Object discovery in videos as foreground motion clustering. In *CVPR*, 2019. 2
- [63] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. In *NeurIPS*, 2022. 1, 2, 4, 7
- [64] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021. 1, 2, 7
- [65] Yanchao Yang, Brian Lai, and Stefano Soatto. Dystab: Unsupervised object segmentation via dynamic-static bootstrapping. In *CVPR*, 2021. 2
- [66] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *CVPR*, 2019. 2

- [67] Nianjin Ye, Chuan Wang, Haoqiang Fan, and Shuaicheng Liu. Motion basis learning for unsupervised deep homography estimation with subspace projection. In *ICCV*, 2021.
2
- [68] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018. 8
- [69] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 3, 4, 8