Supplementary Materials for: Object Discovery in Videos as Foreground Motion Clustering

A. PT-RNN Variants

The three PT-RNN variants to compute the weight $\mathbf{w}_t$, *standard*, *conv*, and *convGRU* are shown in detail in Table 1. For *standard*, we show the equations for a single pixel trajectory. It computes weights based on the pixel embeddings along that trajectory without knowledge of any other trajectories. For *conv*, it uses a 3×3 convolution kernel instead of the standard matrix multiply to include information from neighboring trajectories. Lastly, for *convGRU*, we design this architecture based on the convGRU architecture [1] which has an explicit memory state to capture longer-term dependencies. For all three variants, the hidden state is $\{\mathbf{h}_t, \mathbf{W}_t\}$. However, in the RNN we propagate $\frac{\widetilde{\mathbf{h}}_t}{\widetilde{\mathbf{W}}_t}$, which is the intermediate weighted sum at time t. This allows the network to use knowledge of the previous weights and pixel embeddings to calculate $\mathbf{w}_{t+1}$.

B. Proof of Proposition 1

We note the the following:

$$\begin{aligned}
\underset{\|\mathbf{w}\|_2=1}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} d(\mathbf{w}, \mathbf{y}_i) &= \underset{\|\mathbf{w}\|_2=1}{\operatorname{argmin}} \ \frac{1}{2n}\sum_{i=1}^{n}(1 - \mathbf{w}^\intercal \mathbf{y}_i) \\
&= \underset{\|\mathbf{w}\|_2=1}{\operatorname{argmin}} \left[1 - \frac{1}{n}\sum_{i=1}^{n} \mathbf{w}^\intercal \mathbf{y}_i\right] \\
&= \underset{\|\mathbf{w}\|_2=1}{\operatorname{argmax}} \sum_{i=1}^{n} \mathbf{w}^\intercal \mathbf{y}_i \\
&= \underset{\|\mathbf{w}\|_2=1}{\operatorname{argmax}} \ \mathbf{w}^\intercal \sum_{i=1}^{n} \mathbf{y}_i
\end{aligned}$$

Note that the unit vector that maximizes the inner product with a given vector $\mathbf{v}$ is simply the normalized version of $\mathbf{v}$ (if $\mathbf{v} \neq 0$). Thus, the solution to the above problem is $\frac{\sum_{i=1}^{n} \mathbf{y}_i}{\left\|\sum_{i=1}^{n} \mathbf{y}_i\right\|_2}$.

C. Dataset Details

FT3D The Flying Things 3D dataset (FT3D) [6] is a synthetic dataset comprised of approximately 2250 training and 450 test videos of 10 images each. Each video is created by instantiating a background with static objects and populating the scene by having sampled foreground objects from ShapeNet [5] flying along randomized 3D trajectories. Segmentation masks of all objects (foreground and background) are provided. While [6] does not provide information about which objects are foreground, [11] provided foreground labels by identifying the objects which underwent changes in 3D coordinates. We combined this with the object segmentation masks to produce foreground motion clustering masks. We use this dataset for both evaluation and pre-training. Performance on this dataset is measured by intersection over union (IoU) of the foreground masks.

DAVIS The DAVIS2016 dataset [9] is a collection of 50 videos of approximately 3500 images, split into a 30 training videos and 20 test videos. Each video is accompanied by pixel-dense foreground labels at each frame. We evaluate on the test set for video foreground segmentation only. The DAVIS2017 dataset [10] expands on DAVIS2016 and provides 90 publicly available video sequences with full pixel-dense annotation. DAVIS2017 focuses on semi-supervised video segmentation (as opposed to unsupervised, i.e. foreground segmentation) and provides multiple labels per video. However, not every object labeled is foreground, and not every foreground object is labeled, thus this dataset is not suitable for the task of object discovery. Despite this, we leverage the sequences for training. We use the $\mathcal{J}$-measure (IoU) and the $\mathcal{F}$-measure as defined by [9] as evaluation metrics for DAVIS2016.

FBMS The Freiburg-Berkeley motion segmentation dataset [8] consists of 59 videos split into 29 training videos and 30 test videos. The videos can be up to 800 images long, and approximately every 20th frame has ground truth motion segmentation labels. The inconsistency and ambiguity in motion segmentation dataset labels inspired [2] to rigorously define the problem of motion segmentation and provide corrected labels which we use in this work. Performance on this dataset is measured by precision, recall, F-score, and ΔObj as described in [8, 4].

standard	*conv*	*convGRU*
$\mathbf{c}_t^{i,j} = \text{ReLU}\left(W_c \left[\frac{\widetilde{\mathbf{h}}_{t-1}^{i,j}}{\widetilde{\mathbf{W}}_{t-1}^{i,j}} \quad \mathbf{x}_t^{i,j}\right]\right)$ $\mathbf{w}_t^{i,j} = \sigma\left(W_w \mathbf{c}_t^{i,j}\right)$	$\mathbf{c}_t = \text{ReLU}\left(W_c * \left[\frac{\widetilde{\mathbf{h}}_{t-1}}{\widetilde{\mathbf{W}}_{t-1}} \quad \mathbf{x}_t\right]\right)$ $\mathbf{w}_t = \sigma\left(W_w * \mathbf{c}_t\right)$	$\mathbf{z}_t = \sigma\left(W_z * \left[\frac{\widetilde{\mathbf{h}}_{t-1}}{\widetilde{\mathbf{W}}_{t-1}} \quad \mathbf{x}_t\right]\right)$ $\mathbf{r}_t = \sigma\left(W_r * \left[\frac{\widetilde{\mathbf{h}}_{t-1}}{\widetilde{\mathbf{W}}_{t-1}} \quad \mathbf{x}_t\right]\right)$ $\hat{\mathbf{c}}_t = \text{ReLU}\left(W_{\hat{c}} * \left[\mathbf{r}_t \odot \frac{\widetilde{\mathbf{h}}_{t-1}}{\widetilde{\mathbf{W}}_{t-1}} \quad \mathbf{x}_t\right]\right)$ $\mathbf{c}_t = (1 - \mathbf{z}_t) \odot \widetilde{\mathbf{c}}_{t-1} + \mathbf{z}_t \odot \hat{\mathbf{c}}_t$ $\mathbf{w}_t = \sigma\left(W_w * \mathbf{c}_t\right)$
$\mathbf{h}_t = \widetilde{\mathbf{h}}_{t-1} + \mathbf{w}_t \odot \mathbf{x}_t$ $\mathbf{W}_t = \widetilde{\mathbf{W}}_{t-1} + \mathbf{w}_t$		

Table 1: PT-RNN variants. For *standard*, we show the equations for pixel (i, j), while for the others we show equations in terms of the entire $H \times W \times C$ feature map. Note that for *standard*, $W_c \in \mathbb{R}^{1 \times 2C}$, $W_w \in \mathbb{R}^{1 \times C}$, while for *conv* and *convGRU*, $W_c, W_w, W_z, W_r, W_{\hat{c}}$ are 3×3 convolution kernels. $*$ denotes convolution and σ is the sigmoid nonlinearity.

Others We also show results on the Complex Background [7] and Camouflaged Animal [3] datasets. These datasets are small and contain 5 and 9 sequences, respectively. Labels are corrected and provided by [2]. We use the same metrics for evaluation as the FBMS dataset.

D. DAVIS-m

We hand-select 42 videos from the DAVIS2017 [10] train and val datasets (90 videos total) that roughly satisfy the rubric of [2]. We denote this dataset as DAVIS-m, and use it to supplement the small training dataset of FBMS (29 videos). In hand-selecting these videos, we make sure that only (and all of) the foreground objects are labeled, and that the foreground objects are correctly separated into different objects. For example, the video *classic-car* shows two people in a car with a segmentation mask for the car, and separate segmentation masks for the people. This is incredibly difficult for an algorithm to properly segment using motion cues (and does not fit the rubric of [2]), thus is not included in DAVIS-m. The exact videos are given in Table 2, where we show all 42 videos. There are 27 videos that have a single object (i.e. video foreground segmentation) and 15 videos with multiple objects.

E. Object Discovery results on FT3D

To facilitate motion segmentation and object discovery research, we provide our motion segmentation results for the FT3D [6] testset. We provide numbers for the metrics described in [8, 4], namely precision, recall, F-score, and ΔObj for the multi-object and foreground settings. We trained our full model for 150k iterations using the motion segmentation labels we extracted from foreground labels [11] and object segmentation labels [6]. The results are provided in Table 3.

Multi-object	Foreground
boxing-fisheye	*bear*
cat-girl	*bike-packing*
disc-jockey	*blackswan*
dog-gooses	*breakdance-flare*
dogs-jump	*bus*
gold-fish	*car-shadow*
judo	*car-turn*
kid-football	*cows*
loading	*dance-twirl*
night-race	*dog*
pigs	*drift-chicane*
planes-water	*drift-straight*
sheep	*drift-turn*
tuk-tuk	*elephant*
walking	*flamingo*
	goat
	hike
	koala
	libby
	lucia
	mallard-fly
	mallard-water
	parkour
	rallye
	rhino
	rollerblade
	soccerball

Table 2: DAVIS-m videos. The left column shows the 15 multi-object videos (2 or more objects), and the right column shows the 27 single-object videos (i.e. video foreground segmentation).

Multi-object				Foreground		
P	R	F	ΔObj	P	R	F
74.3	75.1	72.9	2.46	96.4	97.7	96.9

Table 3: Results on FT3D

F. About the name "Object Discovery"

Our definition of object discovery is motivated by the robotic application of discovering objects via their motion. Our definition of object discovery is almost identical to that of (multi-object) "motion segmentation" as defined in [2], except that objects should be tracked even when there is no observed flow at certain frames.

References

[1] N. Ballas, L. Yao, C. Pal, and A. Courville. Delving deeper into convolutional networks for learning video representations. In *International Conference on Learning Representations (ICLR)*, 2016. 1

[2] P. Bideau and E. Learned-Miller. A detailed rubric for motion segmentation. *arXiv preprint arXiv:1610.10033*, 2016. 1, 2, 3

[3] P. Bideau and E. Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *European Conference on Computer Vision (ECCV)*, pages 433–449. Springer, 2016. 2

[4] P. Bideau, A. RoyChowdhury, R. R. Menon, and E. Learned-Miller. The best of both worlds: Combining cnns and geometric constraints for hierarchical motion segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[5] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 1

[6] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[7] M. Narayana, A. Hanson, and E. Learned-Miller. Coherent motion segmentation in moving camera videos using optical flow orientations. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 2

[8] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 2014. 1, 2

[9] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[10] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 1, 2

[11] P. Tokmakov, K. Alahari, and C. Schmid. Learning motion patterns in videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2