Self-supervised Video Object Segmentation by Motion Grouping

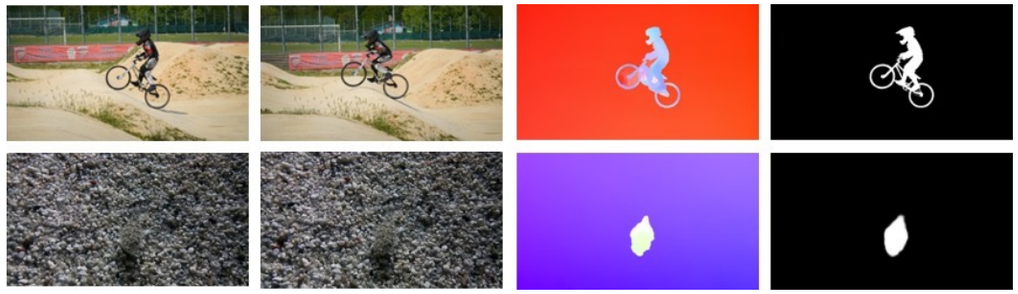
VGG
University of Oxford

Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, Weidi Xie

project page | arxiv | code

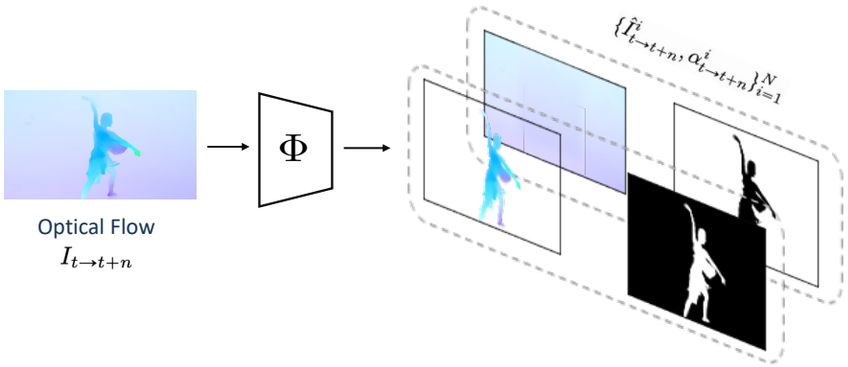
Motivation

- Segmentation is usually carried out using the appearance (RGB) of the frame
- However, objects are not always distinguishable from background. Motion cues can help, e.g. when objects undergo independent motion to the camera
- This work: segmentation using purely motions (no RGB, no annotations)



Intuition

- Layer representation: decompose optical flow to layers
- Each layer represents homogeneous field; discontinuities between layers
- Reconstruction by linear composition of alpha channels (segmentation mask)



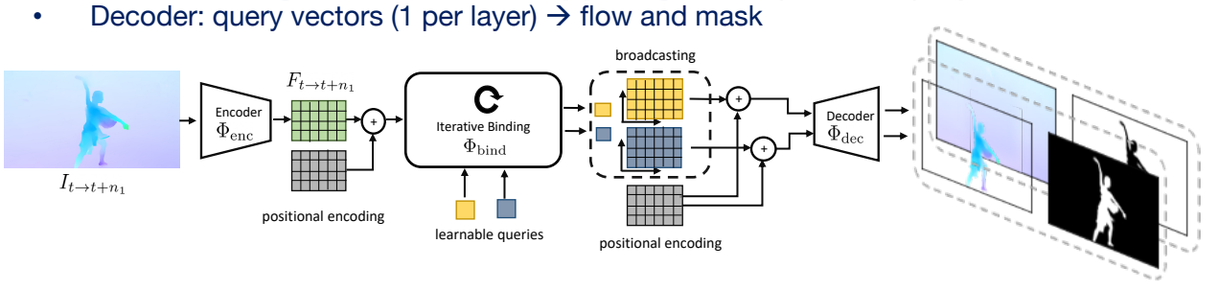
Reconstruction

$$\hat{I}_{t\to t+n} = \sum_{i=1}^{N} \alpha^i_{t\to t+n} \hat{I}^i_{t\to t+n}$$

$$\mathcal{L}_{recon} = \frac{1}{\Omega} \sum_{p\in\Omega} |I_{t\to t+n}(p) - \hat{I}_{t\to t+n}(p)|^2$$

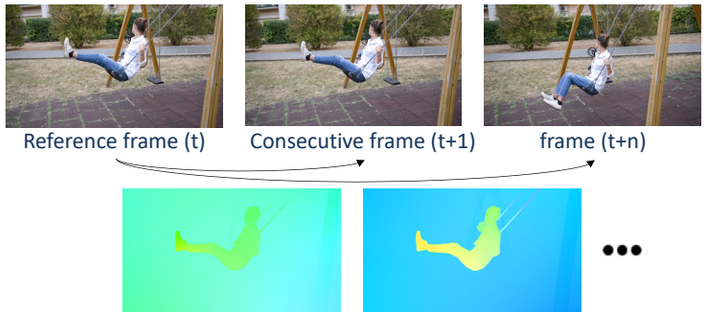
Architecture

- Encoder: input flow → feature maps
- Iterative Binding: Transformer variant [1], assigns each pixel to a query
- Decoder: query vectors (1 per layer) → flow and mask



Temporal Consistency

- Challenge: difficult on objects that undergo similar motion to camera
- Compute flows with different frame gaps, and force mask consistency



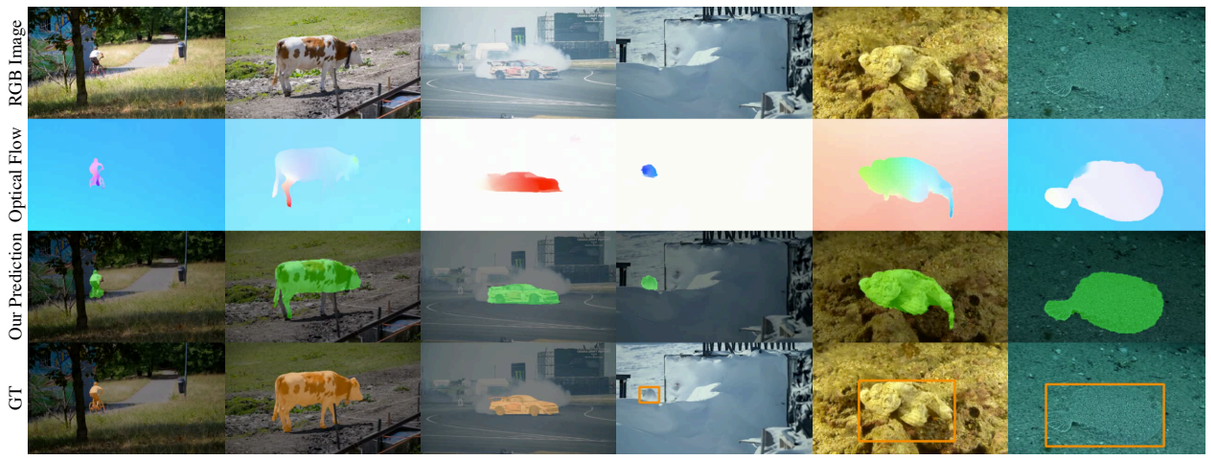
$$\mathcal{L}_{cons} = \frac{1}{\Omega} \min\Big(\sum_{p\in\Omega} |\alpha^1_{t\to t+n_1}(p) - \alpha^1_{t\to t+n_2}(p)|^2, \sum_{p\in\Omega} |\alpha^1_{t\to t+n_1}(p) - \alpha^0_{t\to t+n_2}(p)|^2\Big)$$

Consistency Loss (L2 between masks). Layers are commutative, so loss is **permutation invariant**.

Results

- DAVIS 2016 [2]: the gap between self-supervised and supervised are minimized
- Camouflaged Dataset (MoCA) [3]: competitive with best supervised approaches

Model	Sup.	RGB	Flow	DAVIS16 ($\mathcal{J}$ ↑)	MoCA ($\mathcal{J}$ ↑)	Runtime (s ↓)
CIS (+pp)	✗	✓	✓	59.2 (71.5)	49.4 (54.1)	0.1s (11s)
Ours	✗	✗	✓	**68.3**	**63.4**	**0.012s**
COSNet	✓	✓	✗	80.5	50.7	–
MATNet	✓	✓	✓	**82.4**	**64.2**	0.55s



References

[1] Locatello *et al.* "Object-Centric Learning with Slot Attention". In: NeurIPS2020.
[2] Perazzi *et al.* "A benchmark dataset and evaluation methodology for video object segmentation". In Proc. CVPR, 2016.
[3] Lamdouar *et al.* "Betrayed by motion: Camouflaged object discovery via motion segmentation". In: Proc. ACCV, 2020.