LOCATE: Self-supervised Object Discovery via Flow-guided Graph-cut and Bootstrapped Self-training

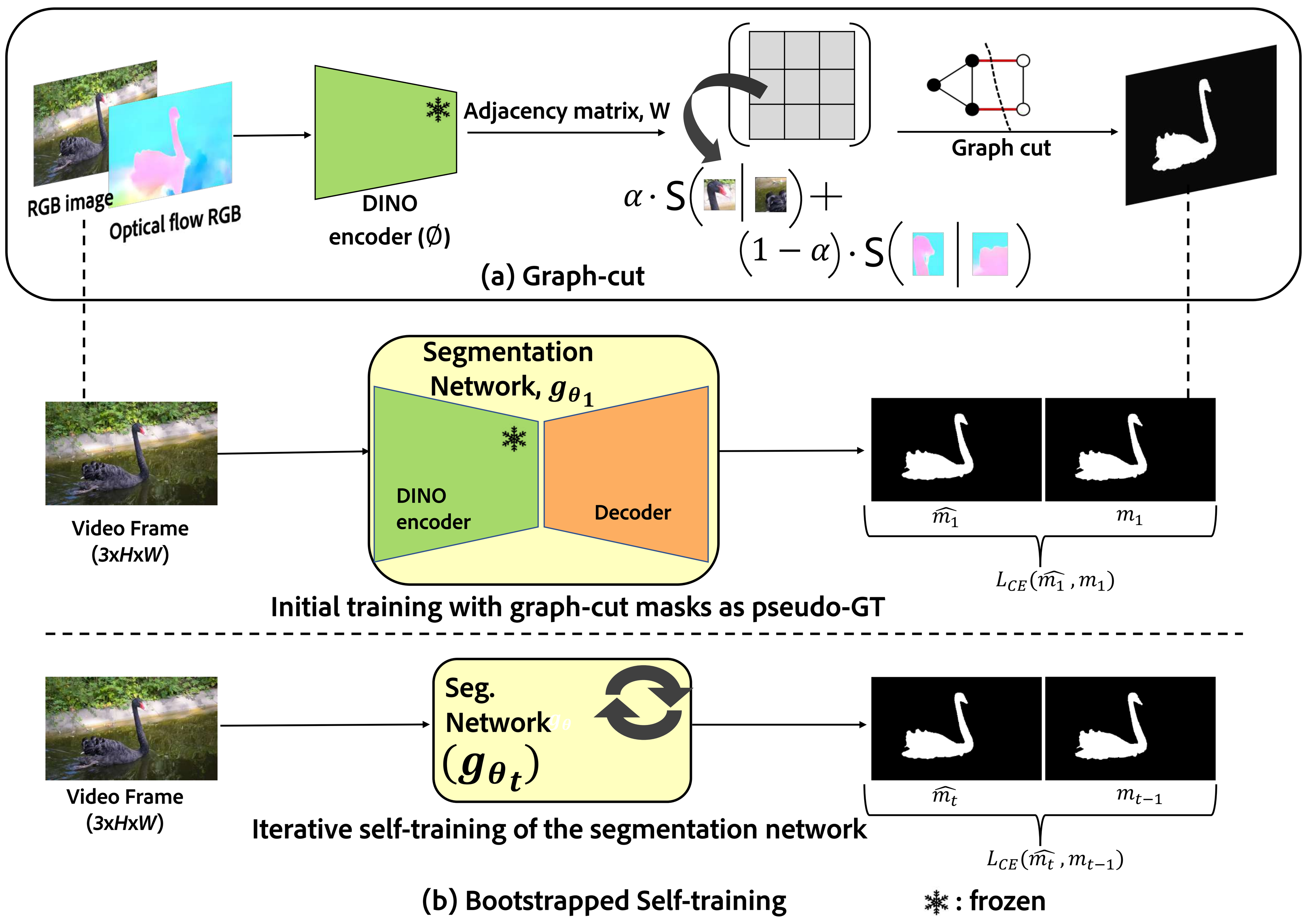
Silky Singh, Shripad Deshmukh, Mausoom Sarkar, Balaji Krishnamurthy
Media and Data Science Research, Adobe

- Code available at: https://github.com/silky1708/LOCATE
- Scan the QR on the right for paper link

1. INTRODUCTION

- self-supervised framework for video object segmentation (VOS)
- matches state-of-the-art (SOTA) performance on DAVIS16
- establishes a new SOTA on SegTrackv2 (**+1 mIoU**)
- inference with single images (no additional inputs required!)
- no post-processing! (deployable in real-world applications)
- trained on videos; exemplary **zero-shot** performance on images

2. METHOD



1. Graph-cut

- Given video frame f, divide f into square patches v_i of size p_s x p_s
- Build a fully-connected graph $G=(V,E)$ on these image patches. $V=\{v_i\}$
- $E(v_i,v_j)$ is given by: cosine similarity (S) scores of the patch features from DINO (∅) [5].
- Specifically,

  $E(v_i,v_j) = \alpha\, S(\emptyset(v_i), \emptyset(v_j)) + (1-\alpha)\, S(\emptyset(flow_i), \emptyset(flow_j))$

  where $\alpha \in [0,1]$, $flow_i$, $flow_j$ are the corresponding optical flow patches.

2. Bootstrapped self-training

- Given N video frames, $x_i \in R^{H\times W\times 3}$, with corresponding graph-cut masks $m_i \in R^{H\times W\times 1}$
- We train a segmentation network, g_θ minimizing cross-entropy(CE):

$$\theta_1^* = \arg\min_{\theta_1} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\mathrm{CE}}(m_i, g_\theta(x_i))$$

- Next, we iteratively train g_θ with its outputs from previous rounds as supervisory signal

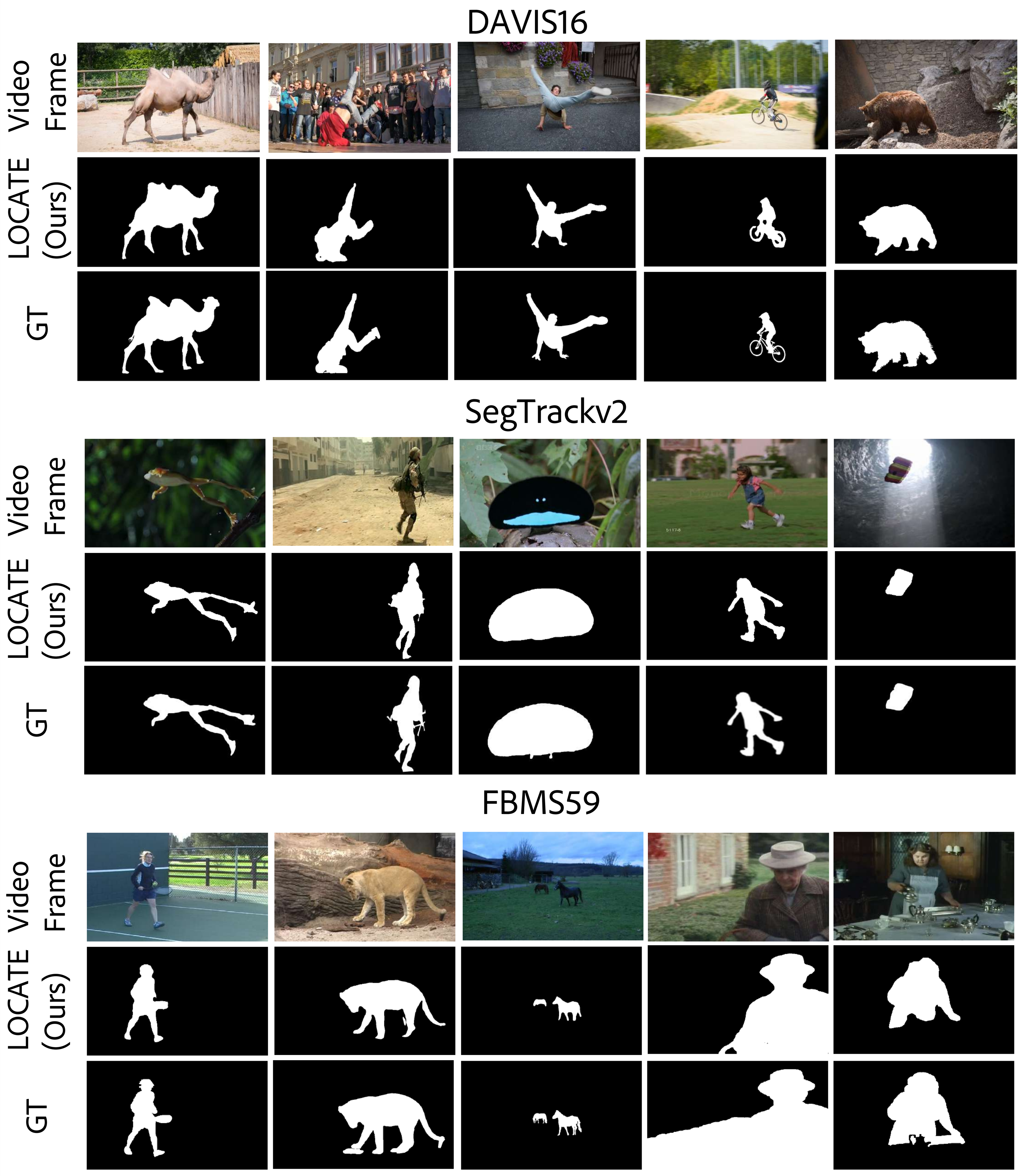
$$\theta_t^* = \arg\min_{\theta_t} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\mathrm{CE}}(g_{\theta_{t-1}^*}(x_i), g_{\theta_t}(x_i))$$

3. QUANTITATIVE RESULTS

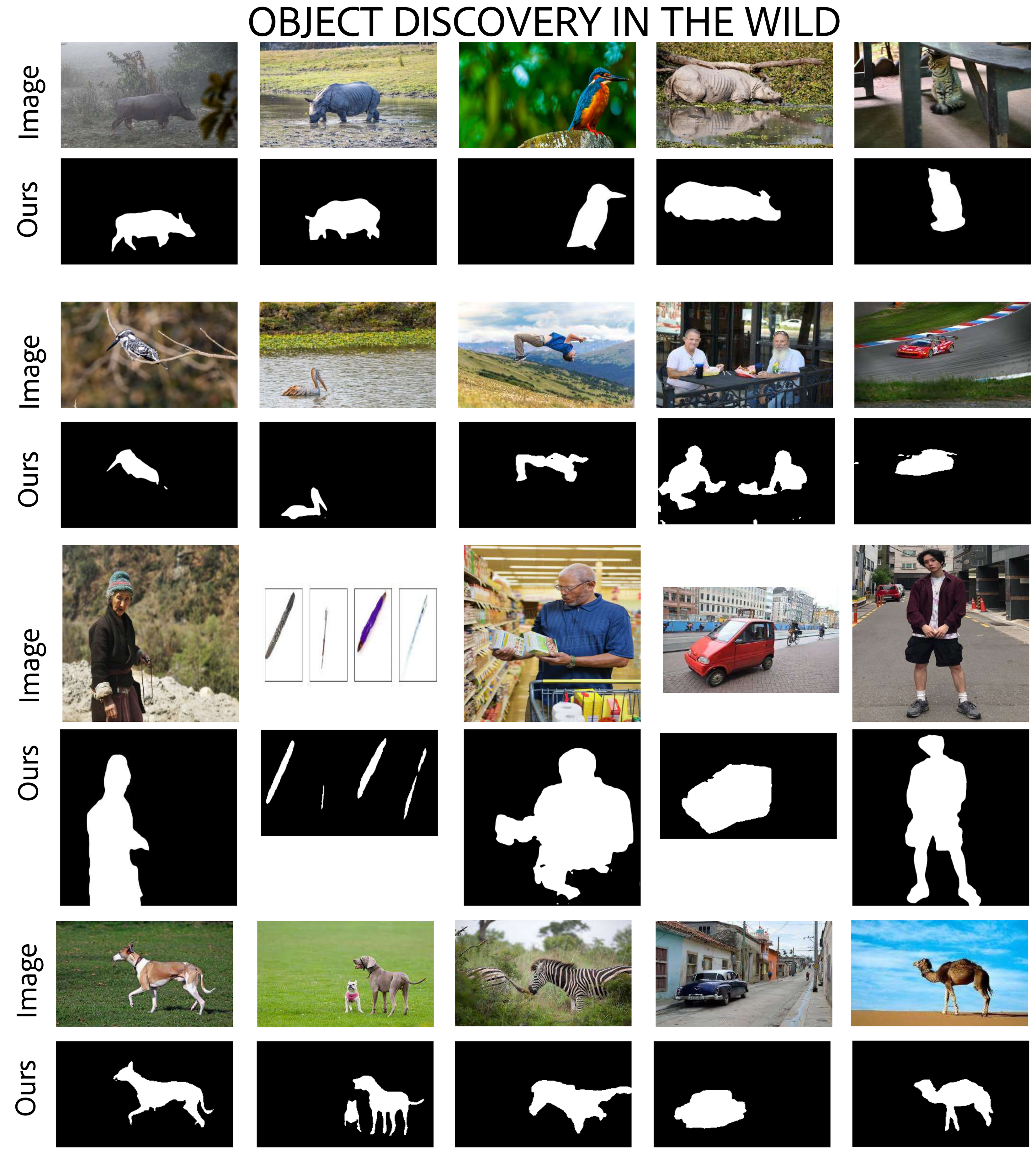
Method	Supervision	Post-processing	DAVIS16 (mIoU)	SegTrackv2 (mIoU)	FBMS59 (mIoU)
Ponimatkin et. al. [1]	None	CRF	80.2	74.9	70.0
OCLR [2]	Synth.	DINO-based TTA	**80.9**	72.3	72.7
DyStaB [3]	Sup. feats.	CRF	80.0	74.2	73.2
GWM [4]	Sup. feats.	CRF + DINO	80.7	78.9	**78.4**
LOCATE (Ours)	None	None	**80.9**	**79.9**	68.8

For detailed comparison, please check out the paper here: https://arxiv.org/abs/2308.11239

4. QUALITATIVE RESULTS



OBJECT DISCOVERY IN THE WILD



5. REFERENCES

[1] Georgy Ponimatkin, Nermin Samet, Yang Xiao, Yuming Du, Renaud Marlet, and Vincent Lepetit. A simple and powerful global optimization for unsupervised video object segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 5892–5903, 2023.
[2] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. In Advances in Neural Information Processing Systems, 2022.
[3] Yanchao Yang, Brian Lai, and Stefano Soatto. Dystab: Unsupervised object segmentation via dynamic-static bootstrapping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2826–2836, 2021.
[4] Subhabrata Choudhury, Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Guess what moves: Unsupervised video and image segmentation by anticipating motion. arXiv preprint arXiv:2205.07844, 2022.
[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9650–9660, 2021.