Guess What Moves:
Unsupervised Video and Image Segmentation by Anticipating Motion

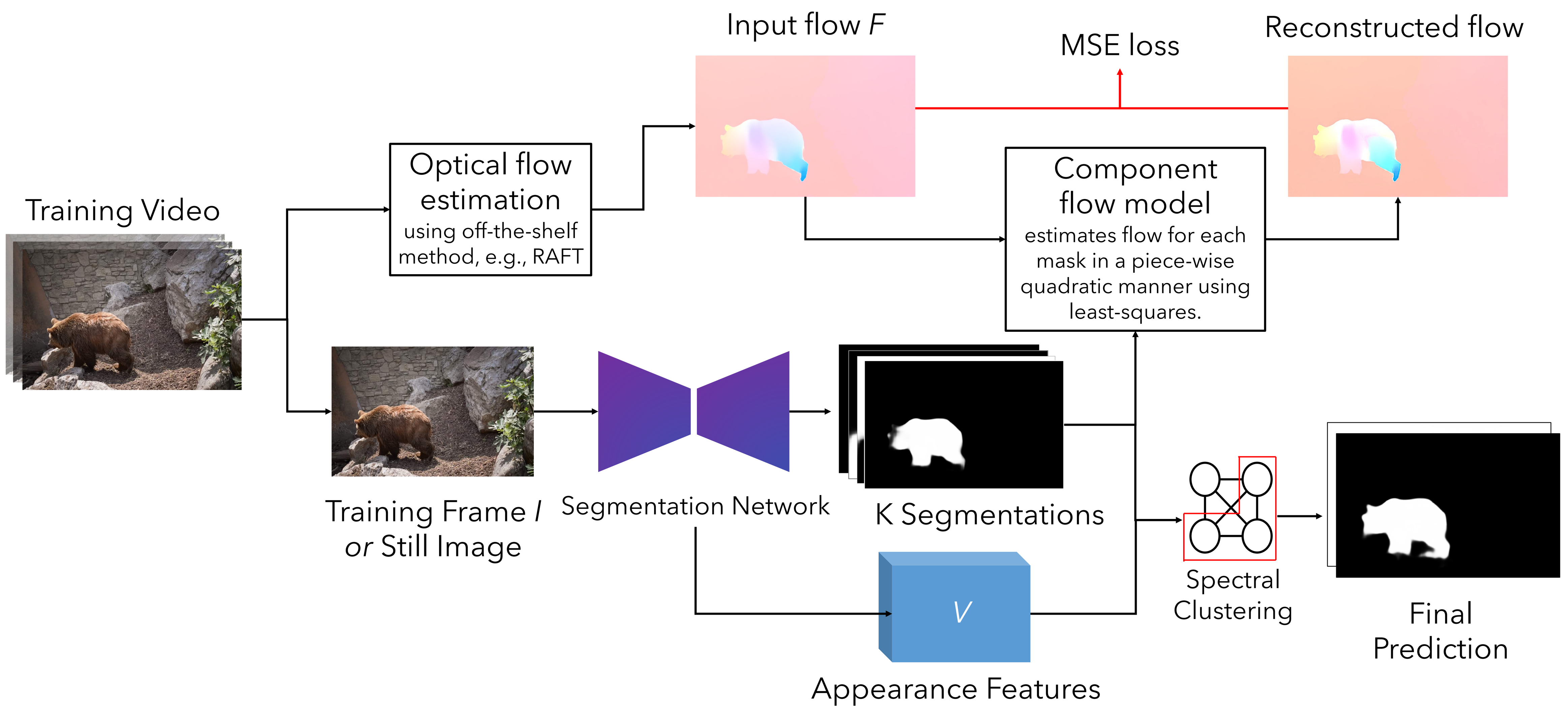
Subhabrata Choudhury*, Laurynas Karazija*, Iro Laina, Andrea Vedaldi, Christian Rupprecht
Visual Geometry Group, University of Oxford

BMVC 2022

Motivation

- **Motion** and more broadly the **principle of common fate** is a useful cue for detecting objects. Motion helps to focus on the parts that might be of interest.
- Some recent unsupervised video segmentation methods **propose using only the motion**, as captured by optical flow, arguing that it provides sufficient information and is easier to model.
- But appearance information can provide **complementary information** to motion cues.
- It can help detect salient objects in absence of motion, in extreme case, detection in still images.
- We instead take a moderate view and re-emphasize the importance of using both appearance and motion modality.
- This enables us to also perform unsupervised image segmentation

Approach

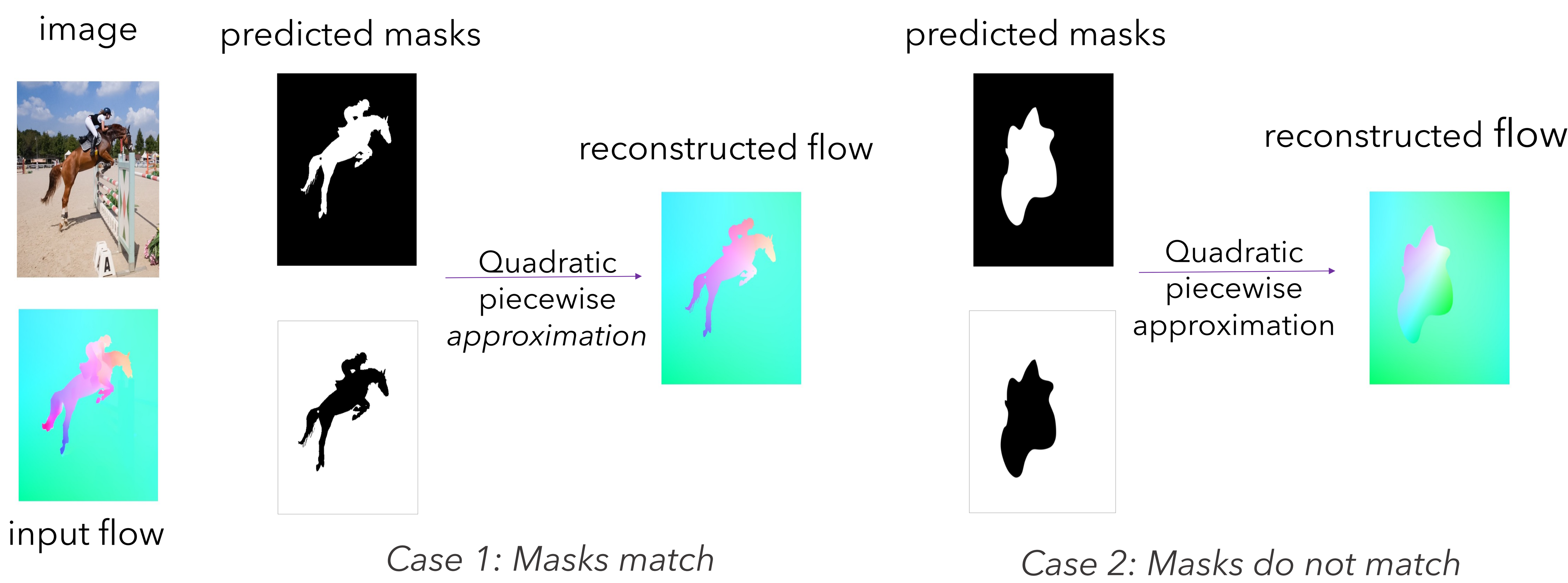


Contribution

- We propose a new **self-supervised approach** that encourages the model to learn a salient object detector from motion cues
- In addition to video object segmentation **we can also perform image segmentation** using the same model without additional training
- We propose a **single network** that can be trained **end-to-end**
- Our method that **does not need motion input during inference** but **can optionally use it**. If motion present we use the full model and train on the frame and flow sets. If motion input is not available we directly use the segmentation network to predict segments
- There it can be applied to videos and still images alike. We test our method on unsupervised **video and image segmentation benchmarks** and achieve comparable results to state-of-the-art methods

Key Idea

- For the flow patterns to be explained well by our simple quadratic model, the masks should roughly correspond to the independent objects
- If the pixels that move together are not grouped together, the component flow model cannot not reconstruct the motion of the scene well



Case 1: Masks match independent objects

Case 2: Masks do not match independent objects

Results

- Segmentation model: MaskFormer with frozen DINO backbone
- Number of segments (K) = 4
- Optical flow method: RAFT

Unsupervised Video Segmentation:

- Datasets: DAVIS, SegTrack v2, FBMS
- Evaluation Metric: Jaccard Index (J)
- Run Mode: Full model

Method	Inf. RGB	Input Flow	Input Resolution	Flow Method	DAVIS J ↑	STv2 J ↑	FBMS J ↑
AMD	✓	✗	128 × 224	-	57.8	57	47.5
MG	✗	✓	128 × 224	RAFT	68.3	58.6	53.1
EM	✗	✓	128 × 224	RAFT	69.3	55.5	57.8
OCLR	✓	✓	480 × 832	RAFT	78.9	71.6	68.7
DS*	✓	✓	240 × 426	RAFT	79.1	72.1	71.8
Ours (UNet)	✓	✗	128 × 224	RAFT	78.3	76.8	72
Ours (Maskformer)	✓	✗	128 × 224	RAFT	**79.5**	**78.3**	**77.4**
CIS*	✓	✓	192 × 384	PWCNet	71.5	62	63.5
DyStaB*	✓	✓	192 × 384	RAFT	80.0	74.2	73.2
Ours* (w/ CRF)	✓	✗	128 × 224	RAFT	**80.7**	**78.9**	**78.4**

- Flow Component model: $F_u \approx A_u + b$
  - $u = [x, x^2, y, y^2, xy] \in R^5$ includes quadratic and mixed terms of the pixel coordinates

Unsupervised Image Segmentation:

- Datasets: CUB, DUTS, ECSSD, DUT-OMRON
- Evaluation Metric: Accuracy (Acc), Jaccard Index (J), F-score
- Run Mode: Only the trained segmentation network is used

	CUB			DUTS			ECSSD			OMRON		
	Acc	J ↑	maxFβ ↑	Acc	J ↑	Fβ ↑	Acc	J ↑	Fβ ↑	Acc	J ↑	Fβ ↑
Voynov et al.	94.0	71.0	80.7	88.1	51.1	60.0	90.6	68.4	79.0	86.0	46.4	53.3
AMD	--	--	--	--	--	60.2	--	--	--	--	--	--
Kyriazi et al.	92.1	66.4	78.3	89.3	52.8	61.4	91.5	71.3	80.6	88.3	50.9	58.3
Kyriazi et al.	--	76.9	--	--	51.4	--	--	73.3	--	--	56.7	--
DyStaB*	--	--	--	--	--	--	--	--	88.1	--	--	73.9
TokenCut	--	--	--	90.3	57.6	--	91.8	71.2	--	88	53.3	--
SelfMask	--	--	--	92.3	62.6	--	94.4	78.1	--	90.1	58.2	--
Ours	93.5	64.6	80.9	91.5	49.2	65.6	88.5	56.1	74.3	89.3	41.31	56.3