High Temporal Consistency through Semantic Similarity Propagation in Semi-Supervised Video Semantic Segmentation for Autonomous Flight

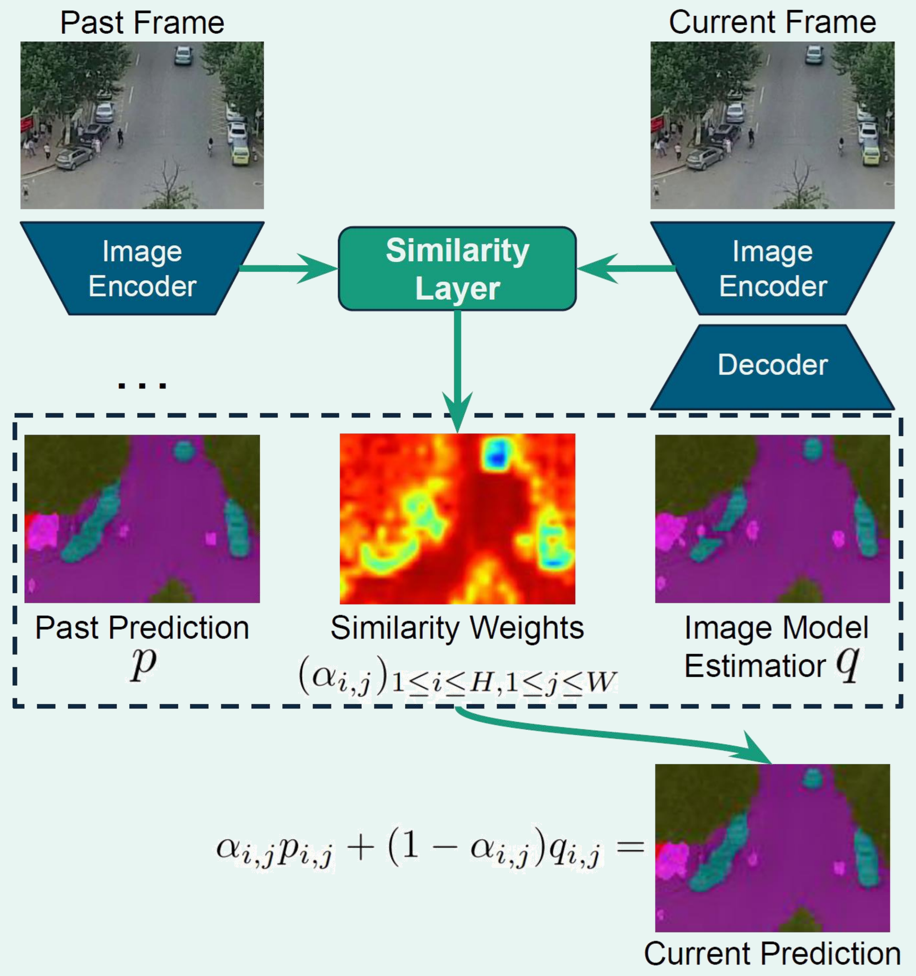
Cédric Vincent[1,2,*], Taehyoung Kim[2,*], Henri Meeß[2]

[1] Télécom Paris, Institut Polytechnique de Paris, [2] Fraunhofer IVI

Introduction

Motivation

Aerial autonomous systems rely on temporally consistent predictions, requiring efficient video segmentation tailored for real-time flight.

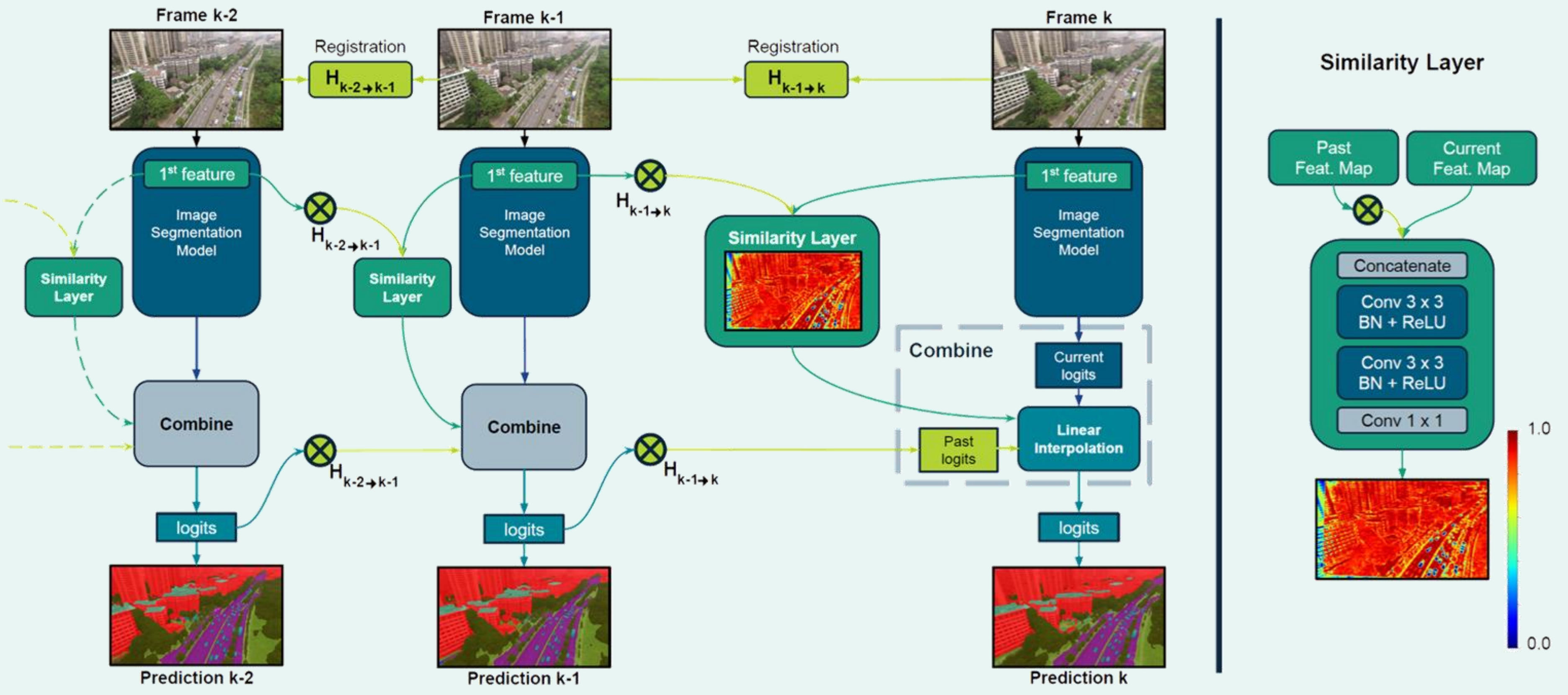


Challenge

Sparse labels and strict inference constraints make it difficult to achieve consistent, accurate segmentation across frames—posing a challenge for safety-critical onboard perception.

Our Approach

We introduce a lightweight propagation method that augments any image segmentation model with temporal consistency via linear interpolation, guided by **learned** semantic similarity and global registration — without sacrificing accuracy or efficiency.



Training

Base Training (SSP)

- We compute the usual cross-entropy segmentation loss on the **labelled current frame *k*.**
- We add an optical flow-based temporal-consistency term between the prediction on *k* and the warped prediction from *k − 1*, down-weighted by a soft occlusion mask $O_{i,j}^{k-1\to k}$.

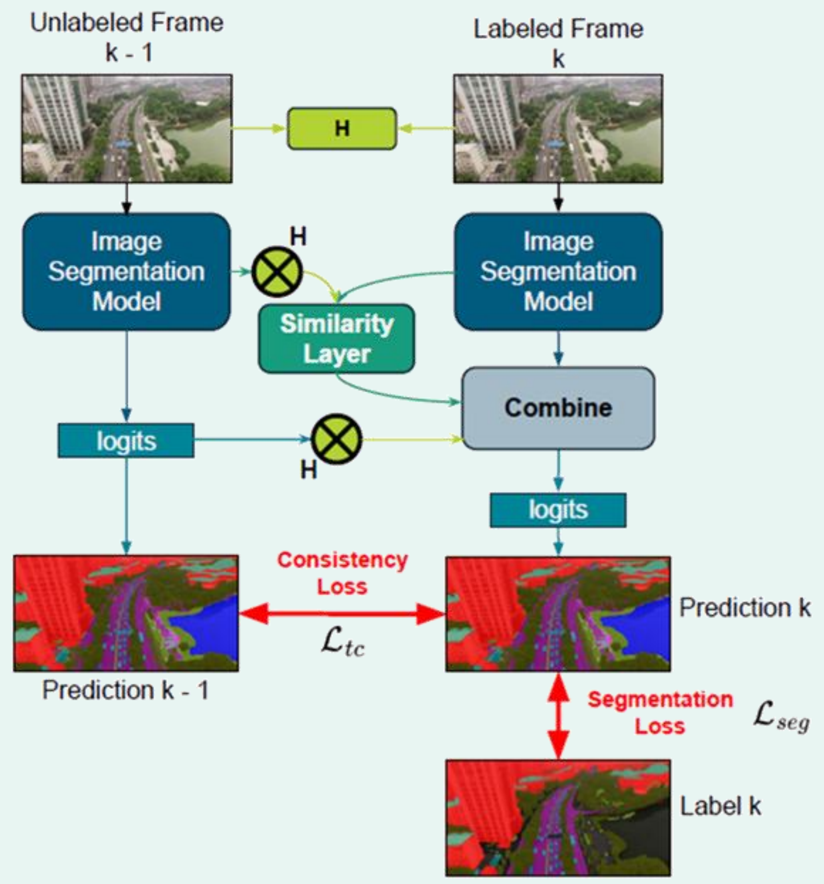
$$\mathcal{L}(P_k, P_{k-1}, A_k) = \underbrace{\mathcal{L}_{\text{seg}}(P_k, A_k)}_{\substack{\text{supervised}\\ \text{segmentation}}} + \lambda\, \underbrace{\mathcal{L}_{\text{tc}}(P_k, P_{k-1})}_{\substack{\text{temporal}\\ \text{consistency}}}$$

$$\mathcal{L}_{\text{tc}} = \frac{1}{HW}\sum_{i,j} O_{i,j}^{k-1\to k}\,\|P_{i,j}^{k} - \hat{P}_{i,j}^{k-1}\|_2^2$$

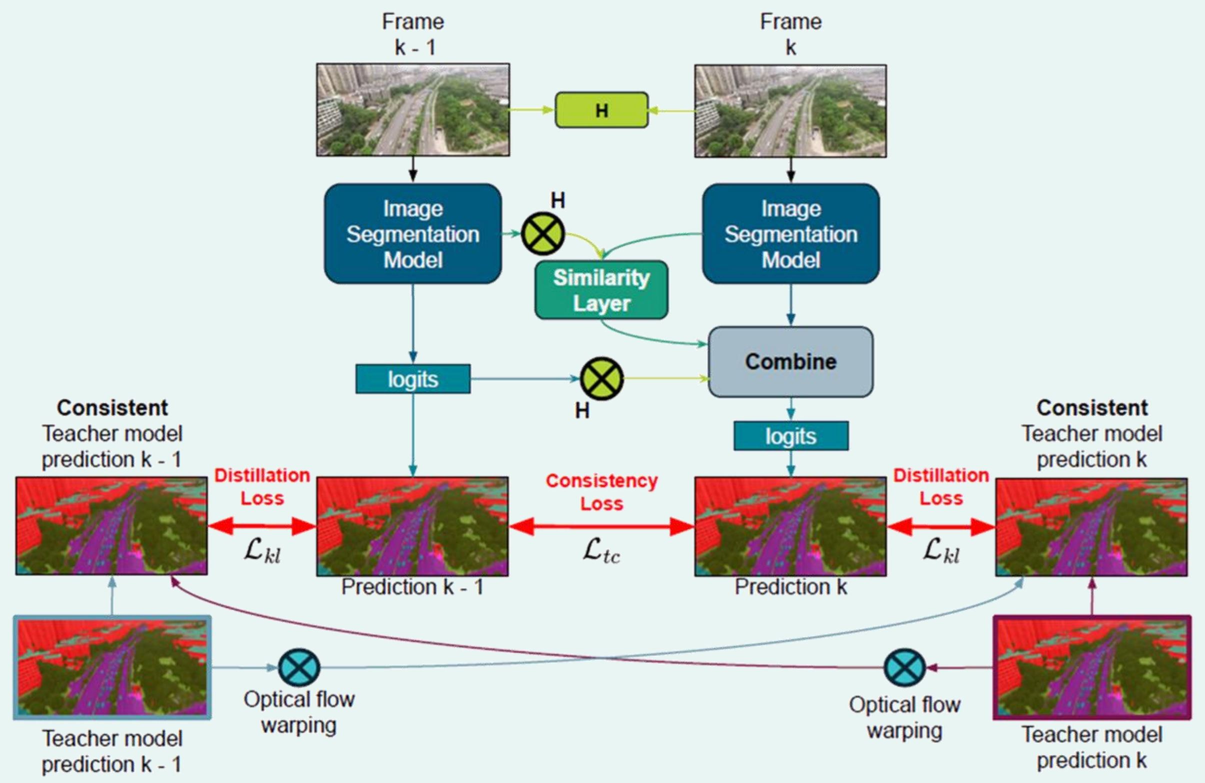
$$O_{i,j}^{k-1\to k} = \exp(-\|I_{i,j}^{k} - \hat{I}_{i,j}^{k-1}\|_1)$$

$P_{i,j}^{k}$: prediction at pixel (i,j) in *current* frame k

$\hat{P}_{i,j}^{k-1}$: prediction from frame $k-1$ warped into frame k



Knowledge Distillation Training (KD-SSP)



- A teacher produces logits for every frame. We warp each set of logits to the opposite frame with bidirectional optical flow and fuse them through an occlusion-mask M_{occ}, yielding **temporally consistent soft labels T_{k-1}^{c} and T_{k}^{c}**
- We supervise the student with KL-divergence loss on both frames, and the optical flow-based temporal-consistency loss ties the two predictions together. Model learn from the all frames while maintaining prediction stability across time.

$$\mathcal{L}(P_k, P_{k-1}, T_k^c, T_{k-1}^c) = \underbrace{\mathcal{L}_{\text{kl}}(P_k, T_k^c)}_{\substack{\text{KD on}\\ \text{current frame } k}} + \underbrace{\mathcal{L}_{\text{kl}}(P_{k-1}, T_{k-1}^c)}_{\substack{\text{KD on}\\ \text{previous frame } k\text{–}1}} + \lambda_{\text{kd}}\ \underbrace{\mathcal{L}_{\text{tc}}(P_k, P_{k-1})}_{\text{temporal consistency}}$$

$$M_{\text{occ}} = \|W_{k\to k-1} + W_{k-1\to k}\|_2^2 > 0.01\left(\|W_{k\to k-1}\|_2^2 + \|W_{k-1\to k}\|_2^2\right) + 0.5$$

$$T_{k-1}^{c} = \frac{T_{k-1} + (1 - M_{\text{occ}})\,(W_{k\to k-1} * T_k)}{2 - M_{\text{occ}}}$$

$$T_{k}^{c} = \frac{T_{k} + (1 - M_{\text{occ}})\,(W_{k-1\to k} * T_{k-1})}{2 - M_{\text{occ}}}$$

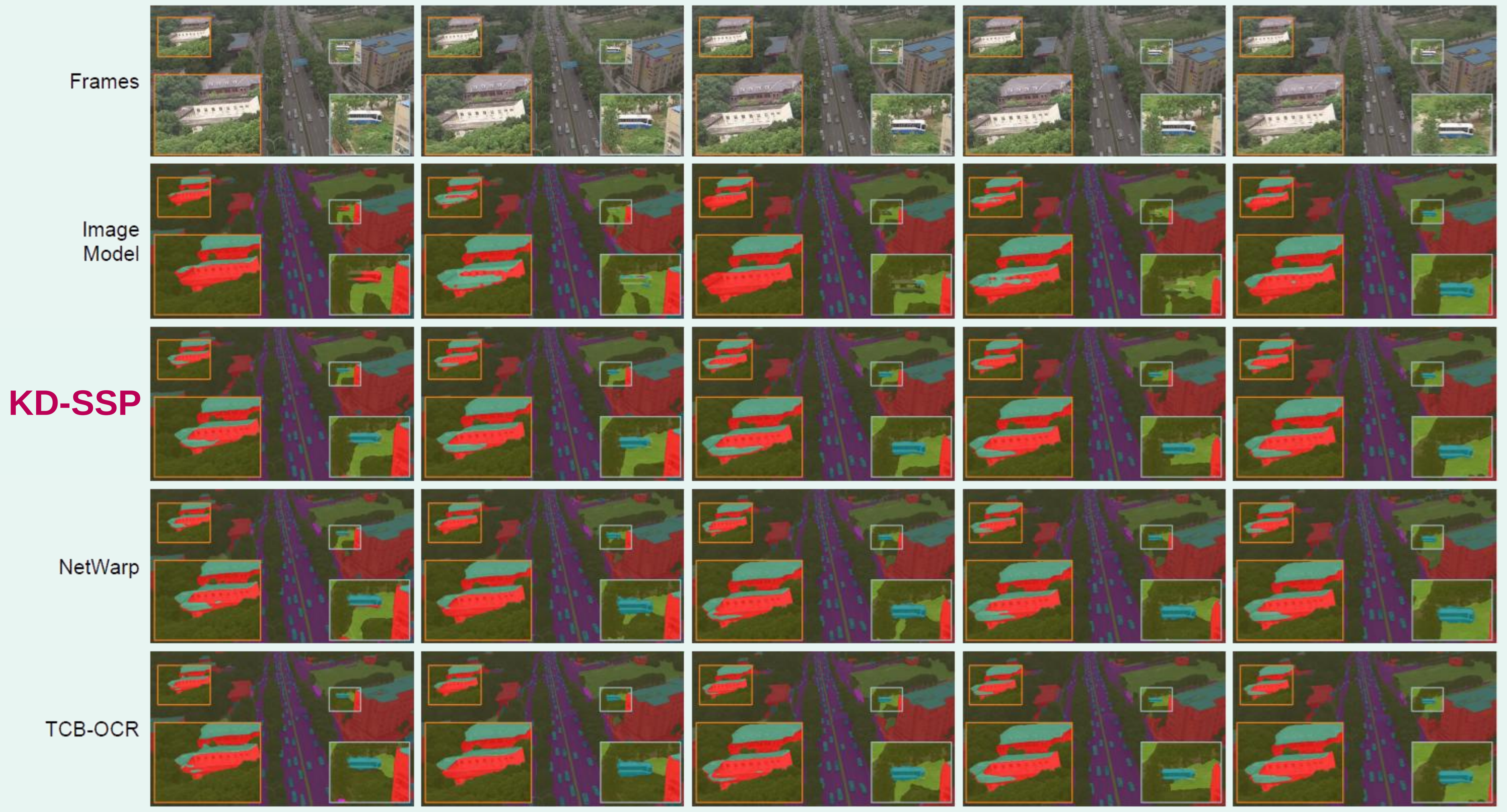
$W_{k-1\to k}$: the optical flow from frame $k-1$ to k

T: the teacher-model prediction on frame k

P: the student-model prediction on frame k

Results

Qualitative Evaluation



Quantitative Evaluation

	Method	Params	GFLOPs	FPS A100	FPS Orin	UAVid mIoU ↑	UAVid TC ↑	RuralScapes mIoU ↑	RuralScapes TC ↑
Image Models	**Teacher Model**	**101.01M**	-	-	-	**81.92**	**84.09**	**66.65**	**89.43**
	SegFormer - b2	**27.36M**	204.1	77	-	77.81	83.76	62.75	86.89
	SegFormer - b3	47.23M	256.7	48	-	78.02	82.59	63.53	86.65
	ConvNeXt-S + UPerNet	81.77M	922.0	96	-	78.35	83.13	63.29	86.70
	Base Image Model	43.17M	310.6	**104**	**31.4**	79.23	79.02	63.51	87.34
	KD Base Image Model (Ours)					80.38	87.15	64.46	90.37
Video Models	DFF [55]	48.43M	**137.2**	23*	-	77.20	83.28	62.66	88.75
	NetWarp [12]	48.44M	739.9	15*	-	79.31	82.19	63.99	88.48
	TCB_{ppm} [33]	64.56M	1350.3	19	-	79.61	81.35	63.83	87.73
	TCB_{ocr} [33]	63.49M	1379.4	18	-	79.67	82.22	63.56	88.39
	SSP (Ours)	43.38M	322.8	95	29.3	79.75	**92.10**	64.00	**94.06**
	KD-SSP (Ours)					**80.63**	91.53	**64.56**	94.00

Contact and further information

Taehyoung Kim
Fraunhofer Institute for Transportation and Infrastructure Systems IVI
taehyoung.kim@ivi.fraunhofer.de

www.ivi.fraunhofer.de/en

Project Page

IVI Home