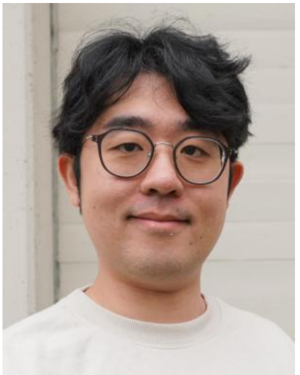
# **SSP**: High Temporal Consistency through **S**emantic **S**imilarity **P**ropagation in Semi-Supervised Video Semantic Segmentation for Autonomous Flight

https://github.com/FraunhoferIVI/SSP


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

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

# Goals

- **Reliable** autonomous systems require **temporally consistent** predictions.
- **Aerial** footage involves **different motion dynamics** than **ground-level** driving footage.

Cityscapes[1]

- Ground-level footage
- Many independently moving objects

UAVid[2]

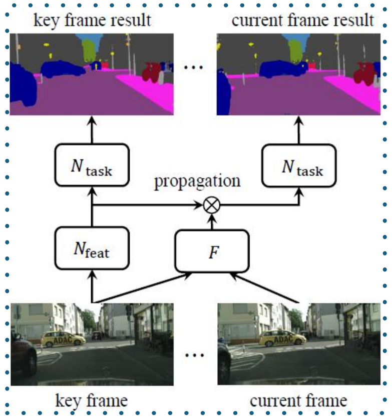
- Aerial footage
- Viewpoint shifts due to UAV motion

[1]M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The Cityscapes Dataset,” in *CVPR Workshop on The Future of Datasets in Vision*, 2015.

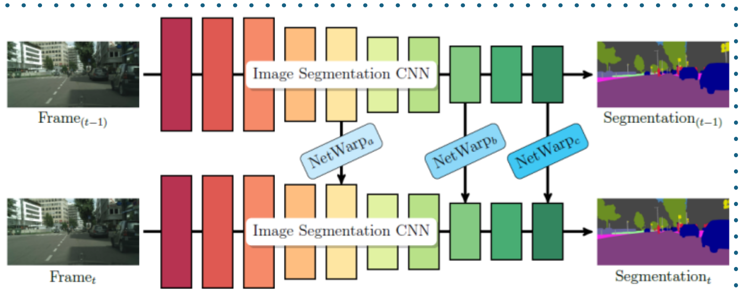
[2]Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 165:108 – 119, 2020

# Challenges

- On-board processing: **Limited computational power** during inference
- **Limited label data**, sparse annotation often not suitable for temporal consistent training

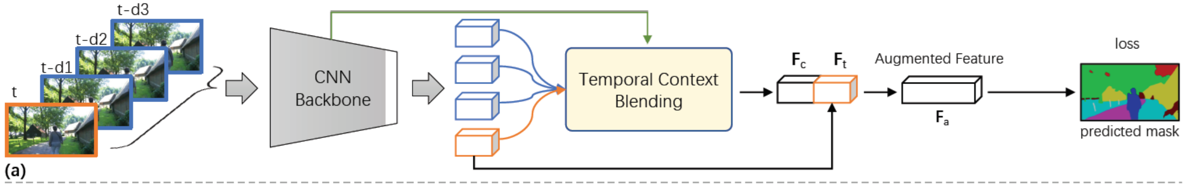



✅ DFF[3] reuses last key-frame features aligned via optical flow for efficient prediction.
❌ Partial computation on non-key frames limits accuracy.




❌ Slow optical flow computation limits real-time performance.

✅ NetWarp[4] Enhances features with temporal context via optical flow and linear interpolation.



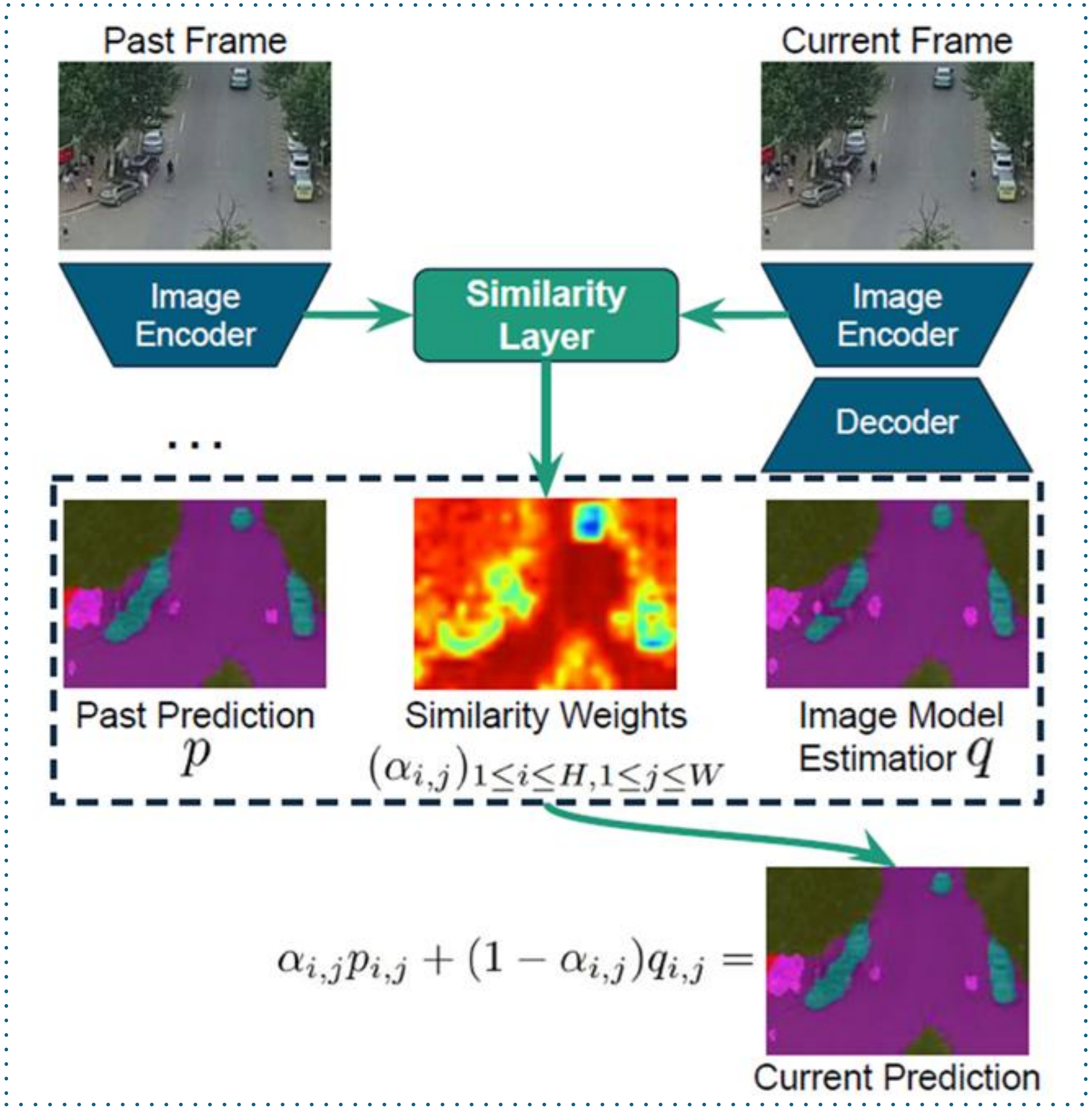

✅ TCB[5] uses attention-based mixing for temporal feature enhancement.
❌ High computation hinders real-time deployment.

✅ SSP focuses on real-time video segmentation tailored for UAVs, balancing accuracy, consistency, and efficiency.
✅ SSP efficiently augments any image semantic segmentation model
✅ KD-SSP leverages unlabeled frames to enhance performance.

[3] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2349–2358, 2017.
[4] Raghudeep Gadde, Varun Jampani, and Peter V. Gehler. Semantic video cnns through representation warping. In The IEEE International Conference on Computer Vision (ICCV), 2017.
[5] Jiaxu Miao, YunchaoWei, YuWu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.

# Contributions

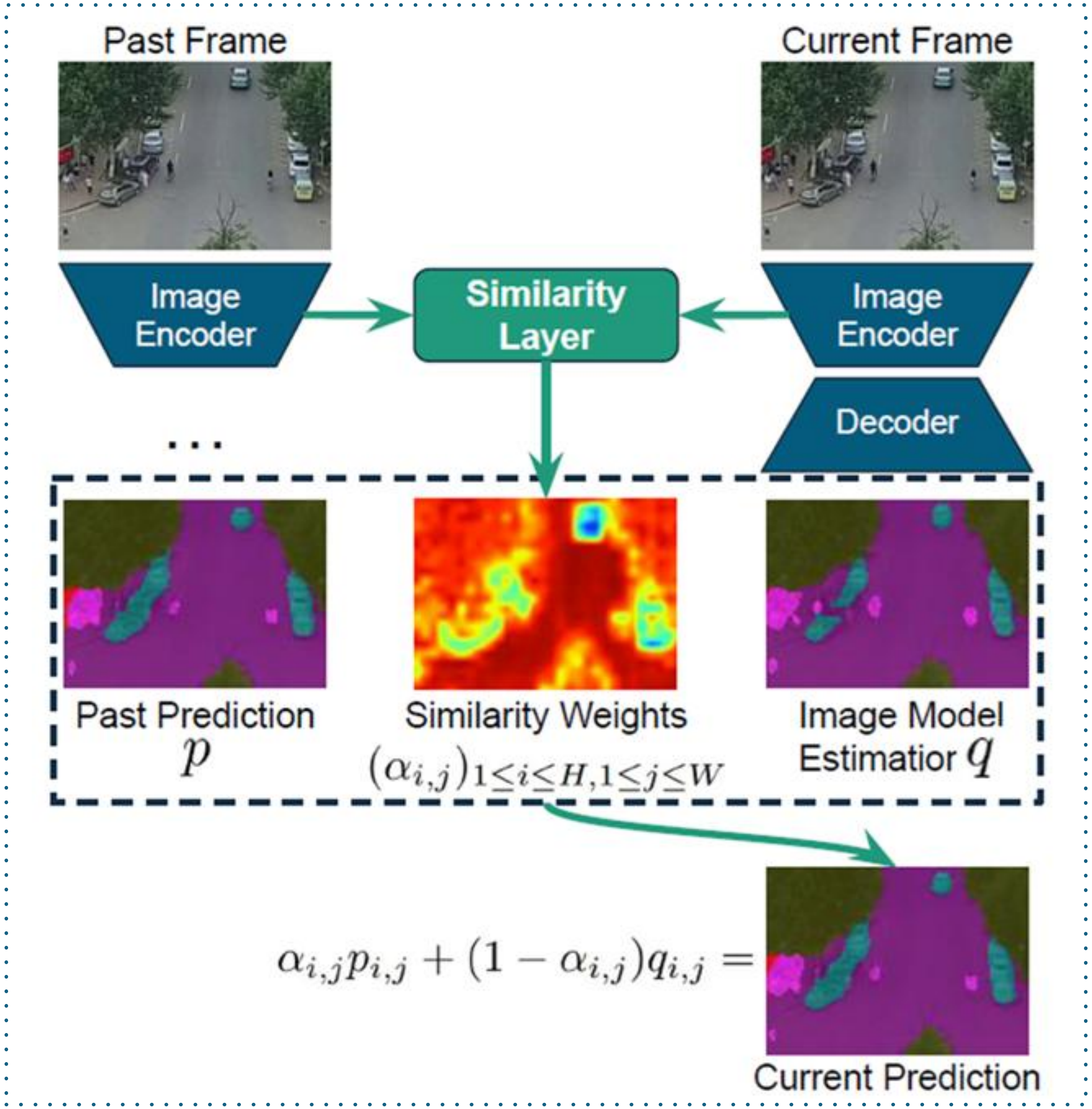
- Simple temporal propagation of predictions



# Contributions

- Simple temporal propagation of predictions
- Global registration alignment

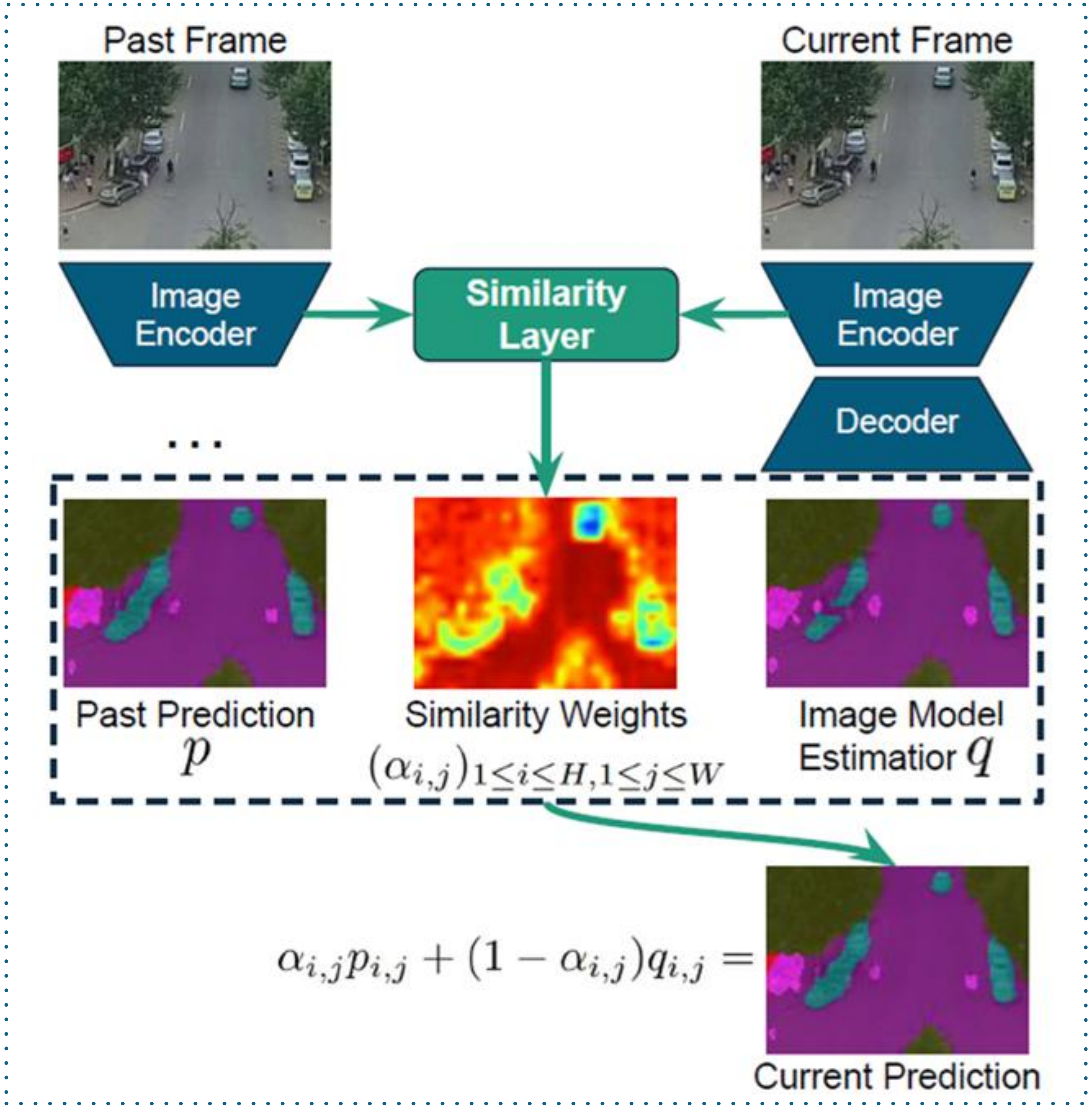
$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$



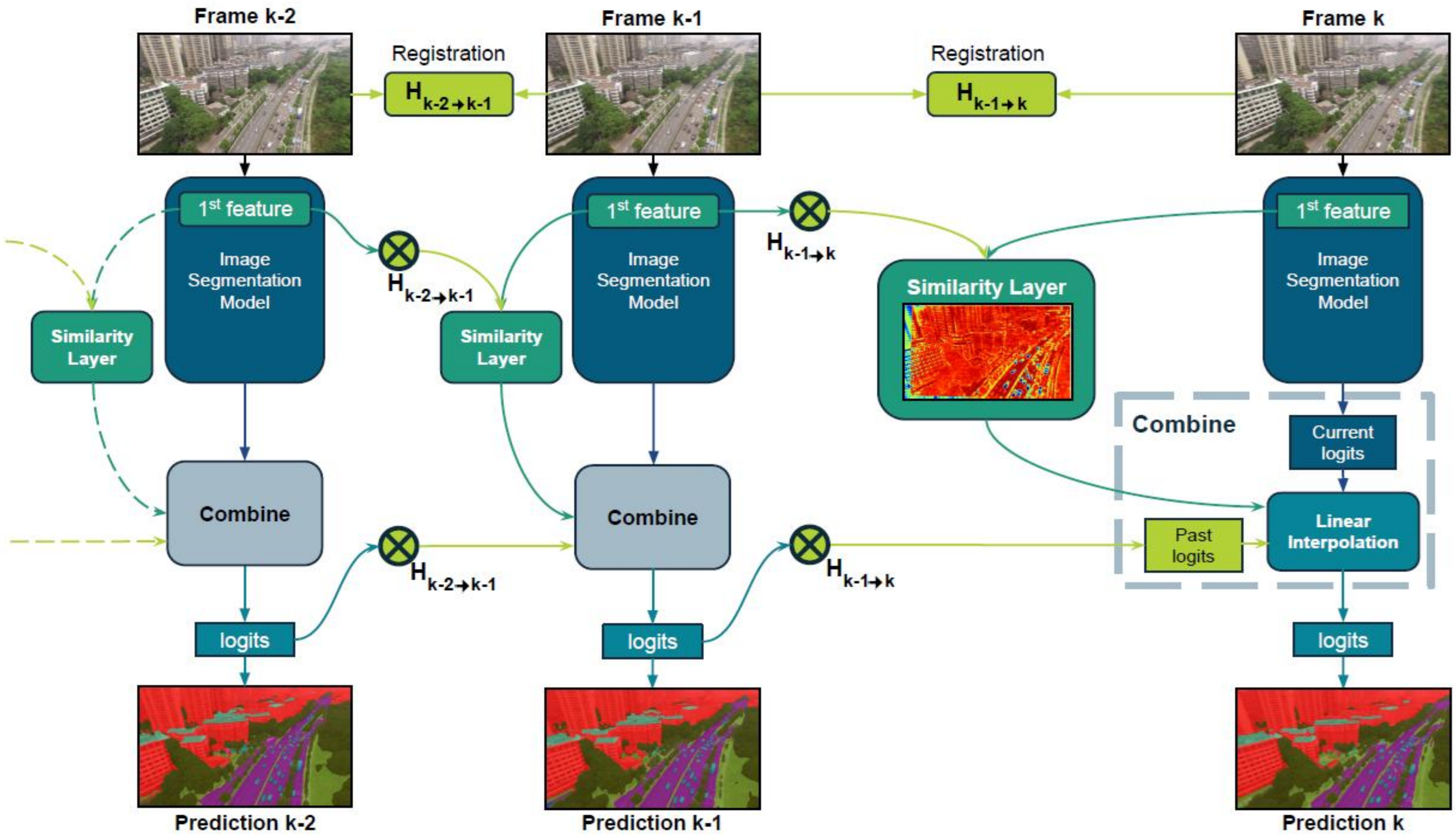
# Contributions

- Simple temporal propagation of predictions
- Global registration alignment
- Consistency-aware knowledge distillation to train efficient models on sparsely labeled datasets

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

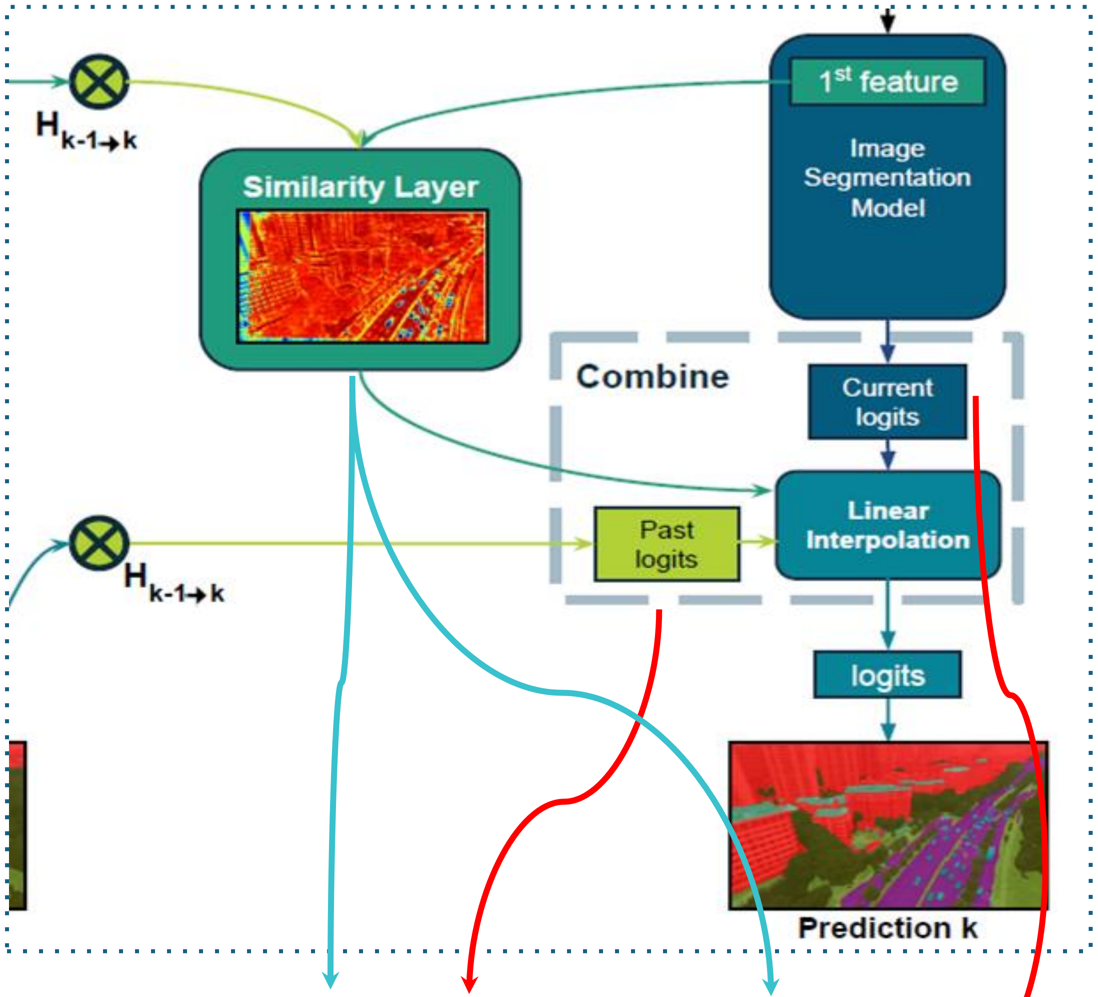


# SSP

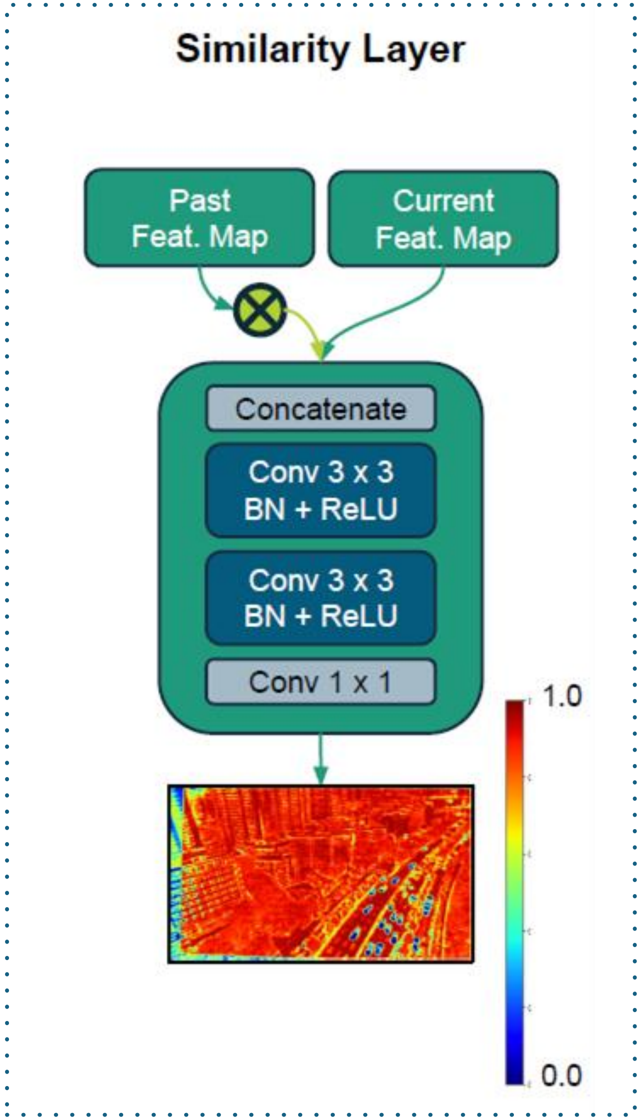


Frame k-2
Frame k-1
Frame k
Registration
Registration
H k-2→k-1
H k-1→k
1st feature
Image Segmentation Model
1st feature
Image Segmentation Model
1st feature
Image Segmentation Model
H k-2→k-1
H k-1→k
Similarity Layer
Similarity Layer
Similarity Layer
Combine
Current logits
Combine
Combine
Past logits
Linear Interpolation
H k-2→k-1
H k-1→k
logits
logits
logits
Prediction k-2
Prediction k-1
Prediction k

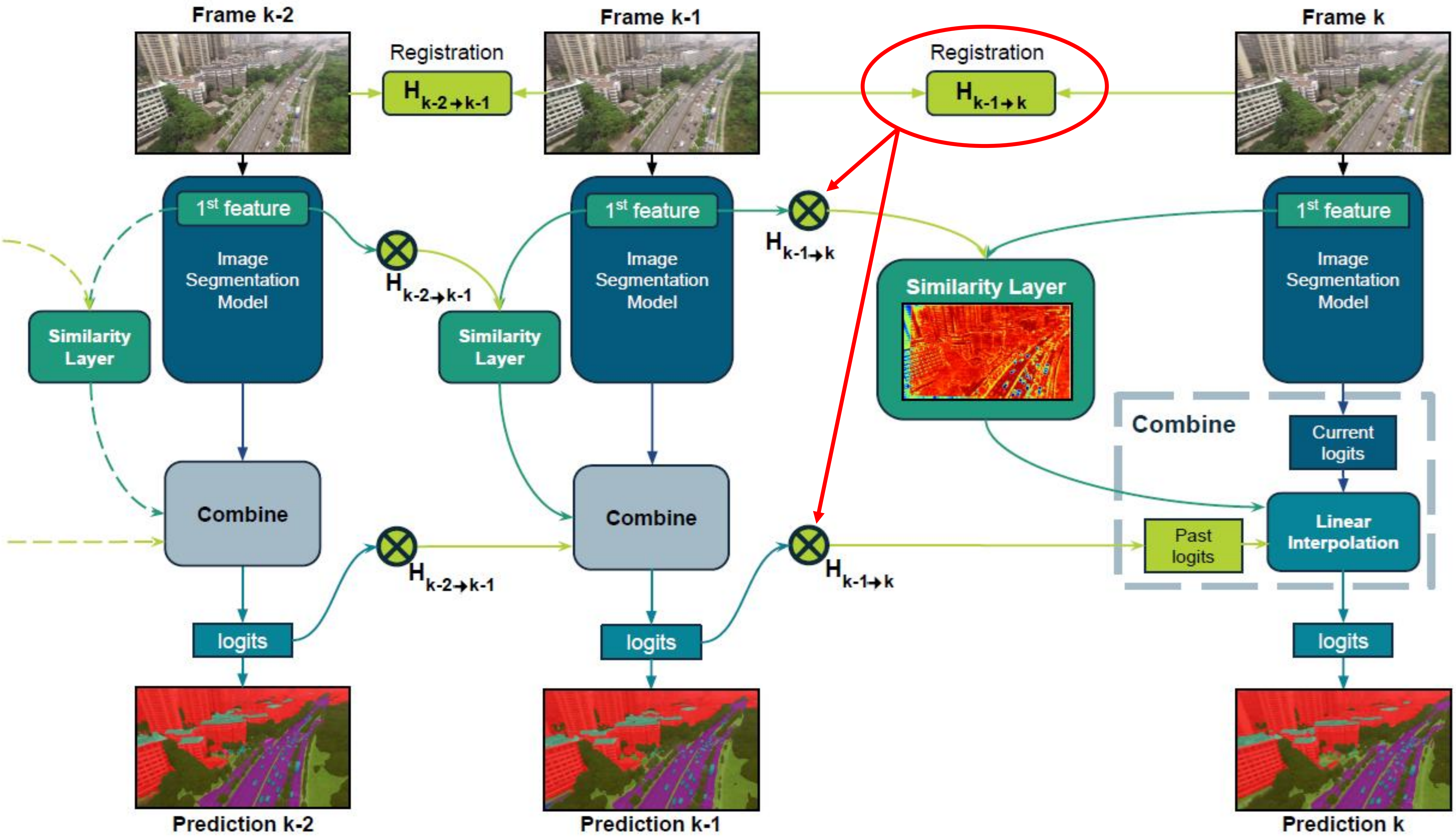
# SSP




$$\hat{q}_{i,j} = \alpha_{i,j} p_{i,j} + (1 - \alpha_{i,j}) q_{i,j}$$



# SSP



Frame k-2
Frame k-1
Frame k
Registration
$H_{k-2 \to k-1}$
Registration
$H_{k-1 \to k}$
1st feature
Image Segmentation Model
1st feature
Image Segmentation Model
1st feature
Image Segmentation Model
$H_{k-2 \to k-1}$
$H_{k-1 \to k}$
Similarity Layer
Similarity Layer
Similarity Layer
Combine
Combine
Combine
Current logits
Linear Interpolation
Past logits
$H_{k-2 \to k-1}$
$H_{k-1 \to k}$
logits
logits
logits
Prediction k-2
Prediction k-1
Prediction k

# SSP

- Without knowledge distillation

$$\mathcal{L}_{tc} = \frac{1}{HW} \sum_{i,j} \mathbf{O}_{i,j}^{k-1 \rightarrow k} \left\| y_{i,j} - \hat{x}_{i,j} \right\|_2^2$$

Current prediction

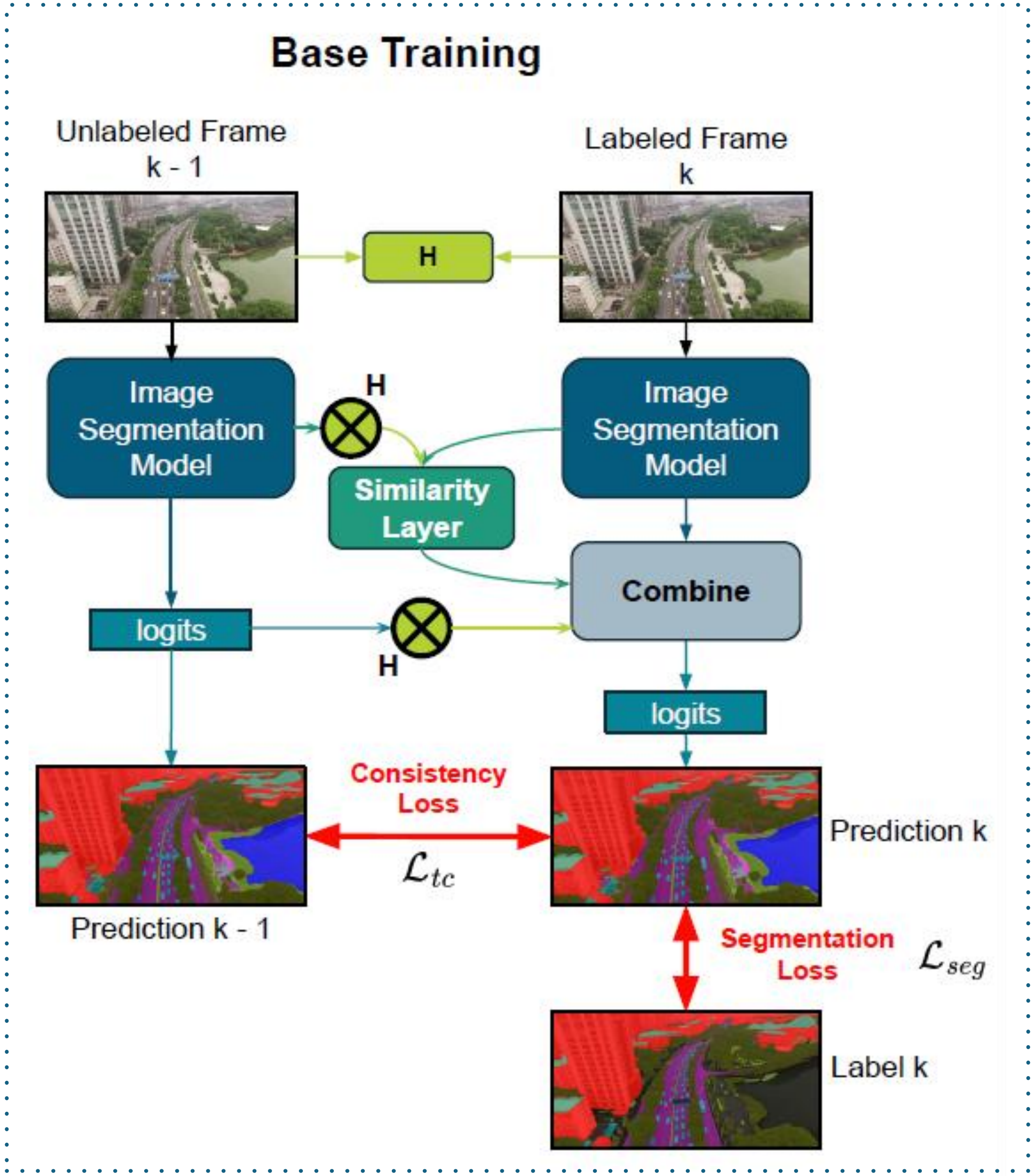
Previous prediction aligned with optical flow

Occlusion Mask:

$$\mathbf{O}_{i,j}^{k-1 \rightarrow k} = exp(- \left\| \mathbf{I}_{i,j}^{k} - \hat{\mathbf{I}}_{i,j}^{k-1} \right\|_1)$$

Current frame

Previous frame aligned with optical flow



# SSP (Occlusion Mask)

- Without knowledge distillation

$$\mathcal{L}_{tc} = \frac{1}{HW} \sum_{i,j} \mathbf{O}_{i,j}^{k-1 \rightarrow k} \left\| y_{i,j} - \hat{x}_{i,j} \right\|_2^2$$

Current prediction

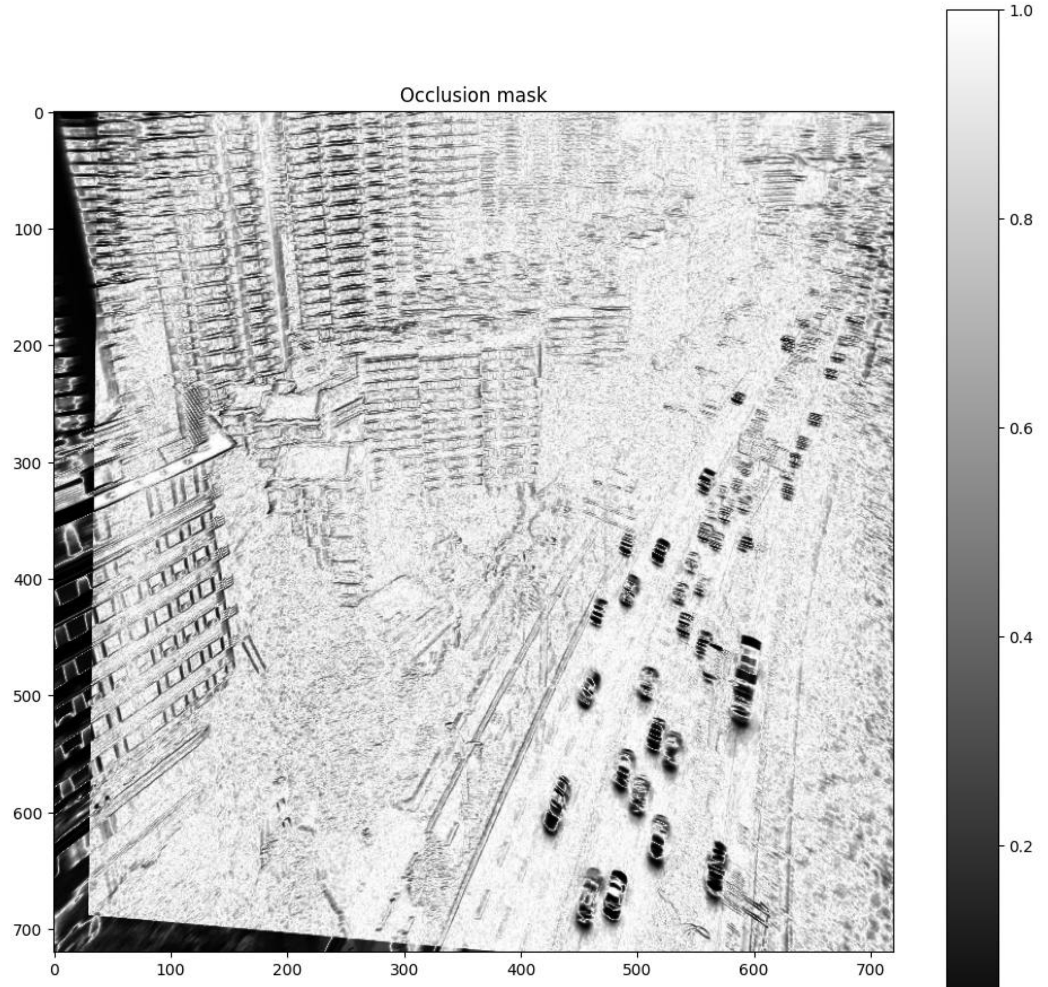
Previous prediction aligned with optical flow

Occlusion Mask:

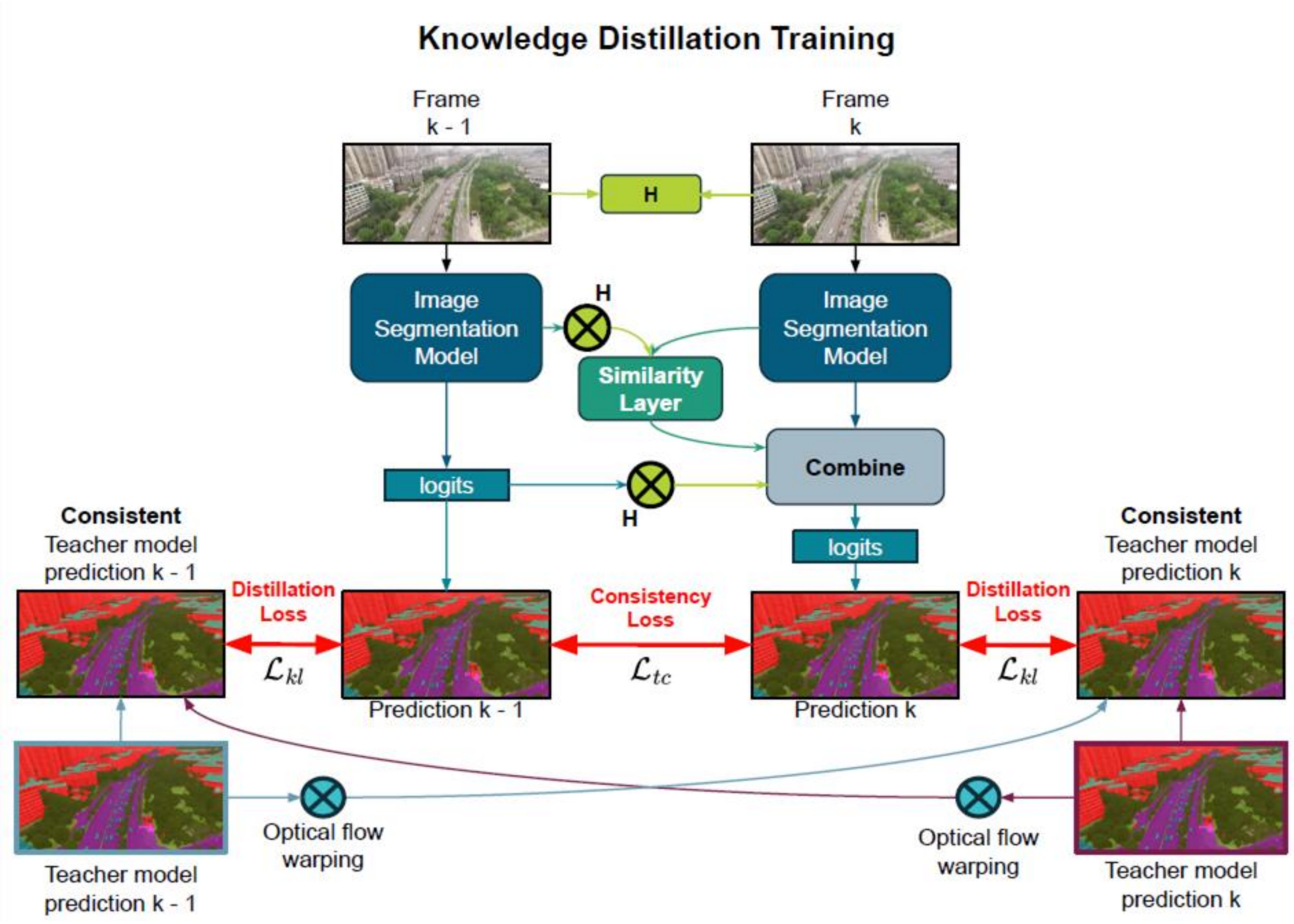
$$\mathbf{O}_{i,j}^{k-1 \rightarrow k} = exp(- \left\| \mathbf{I}_{i,j}^{k} - \hat{\mathbf{I}}_{i,j}^{k-1} \right\|_1)$$

Current frame

Previous frame aligned with optical flow

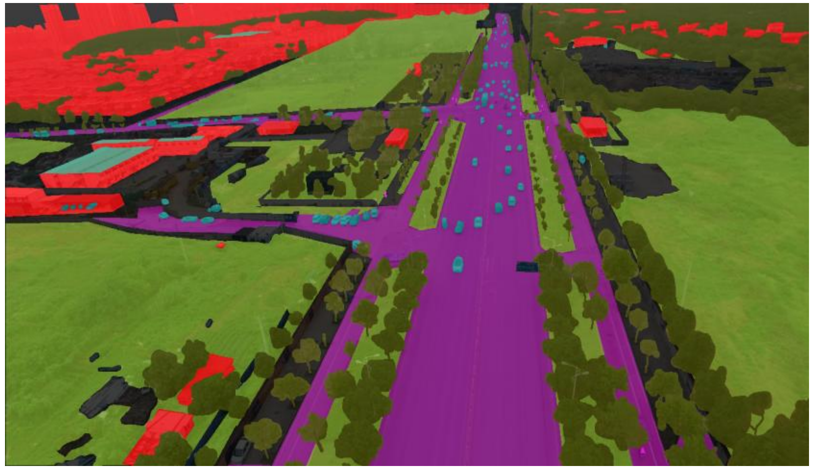


# KD-SSP


Knowledge Distillation Training
Frame k - 1
Frame k
H
Image Segmentation Model
H
Similarity Layer
Image Segmentation Model
Combine
logits
H
logits
Consistent Teacher model prediction k - 1
Distillation Loss
$\mathcal{L}_{kl}$
Consistency Loss
$\mathcal{L}_{tc}$
Distillation Loss
$\mathcal{L}_{kl}$
Consistent Teacher model prediction k
Prediction k - 1
Prediction k
Optical flow warping
Optical flow warping
Teacher model prediction k - 1
Teacher model prediction k

# Experiments

## UAVid[2] Dataset



[2]Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 165:108 – 119, 2020

# Experiments

RuralScapes[6] Dataset

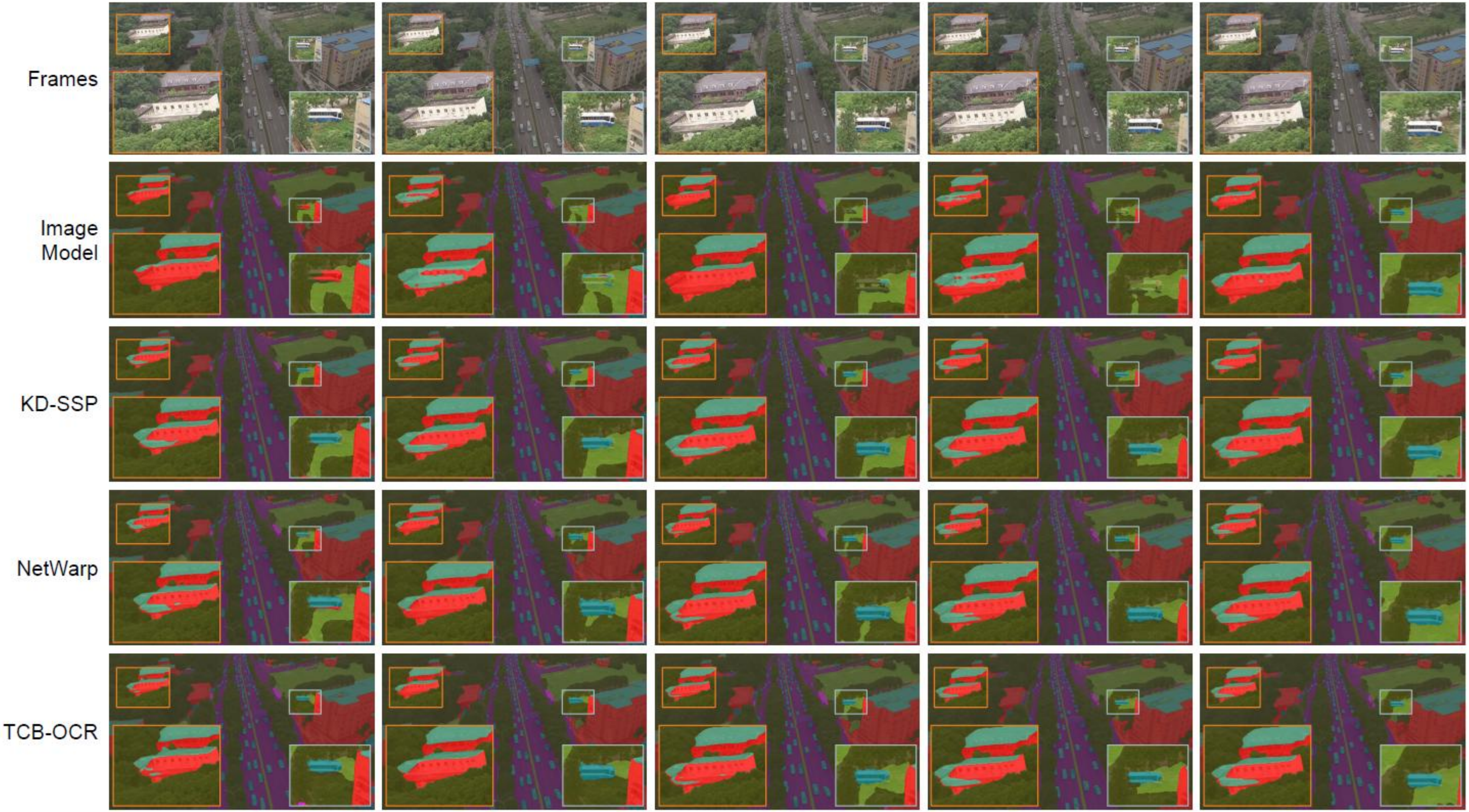
[6]Alina Marcu, Vlad Licaret, Dragos Costea, and Marius Leordeanu. Semantics through time: Semi-supervised segmentation of aerial videos with iterative label propagation. In *Computer Vision – ACCV 2020*, pages 537–552, Cham, 2021. Springer International Publishing

# Results

- Base Image Model: Hiera-S + UPerNet
- KD-SSP achieves superior accuracy-speed trade-off and higher consistency

| | Method | Params | GFLOPs | FPS | | UAVid | | RuralScapes | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | A100 | Orin | mIoU ↑ | TC ↑ | mIoU ↑ | 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 |
| | $\text{TCB}_{ppm}$ [33] | 64.56M | 1350.3 | 19 | - | 79.61 | 81.35 | 63.83 | 87.73 |
| | $\text{TCB}_{ocr}$ [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 |

# Qualitative Results



# Ablation study

- Similarity layer design:
  - *Learned convolutional layers > cosine similarity*
- No global registration alignment:
  - *No loss of accuracy, but lower consistency*
- No propagation of predictions through interpolation:
  - *Same accuracy, lower consistency*
- Consistency loss is essential to higher TC
- Same results without training the base image model on the dataset

| Model | | mIoU | TC |
|---|---|---|---|
| Base image model | | 79.23 | 79.02 |
| SSP | Base parameters | 79.87 | **92.01** |
| | Cos. Sim. Interpolation (TFC) | 79.20 | 91.67 |
| | No registration alignment | 79.84 | 89.65 |
| | No interpolation | 79.99 | 88.20 |
| | No consistency loss | **80.01** | 86.38 |
| | 1-step training | 79.92 | 91.97 |

Table 3. **Ablation study of SSP.** *Base parameters* corresponds to the described SSP configuration used for Tab. 1. Results are obtained without knowledge distillation. UAVid dataset, 736 × 1280 resolution.

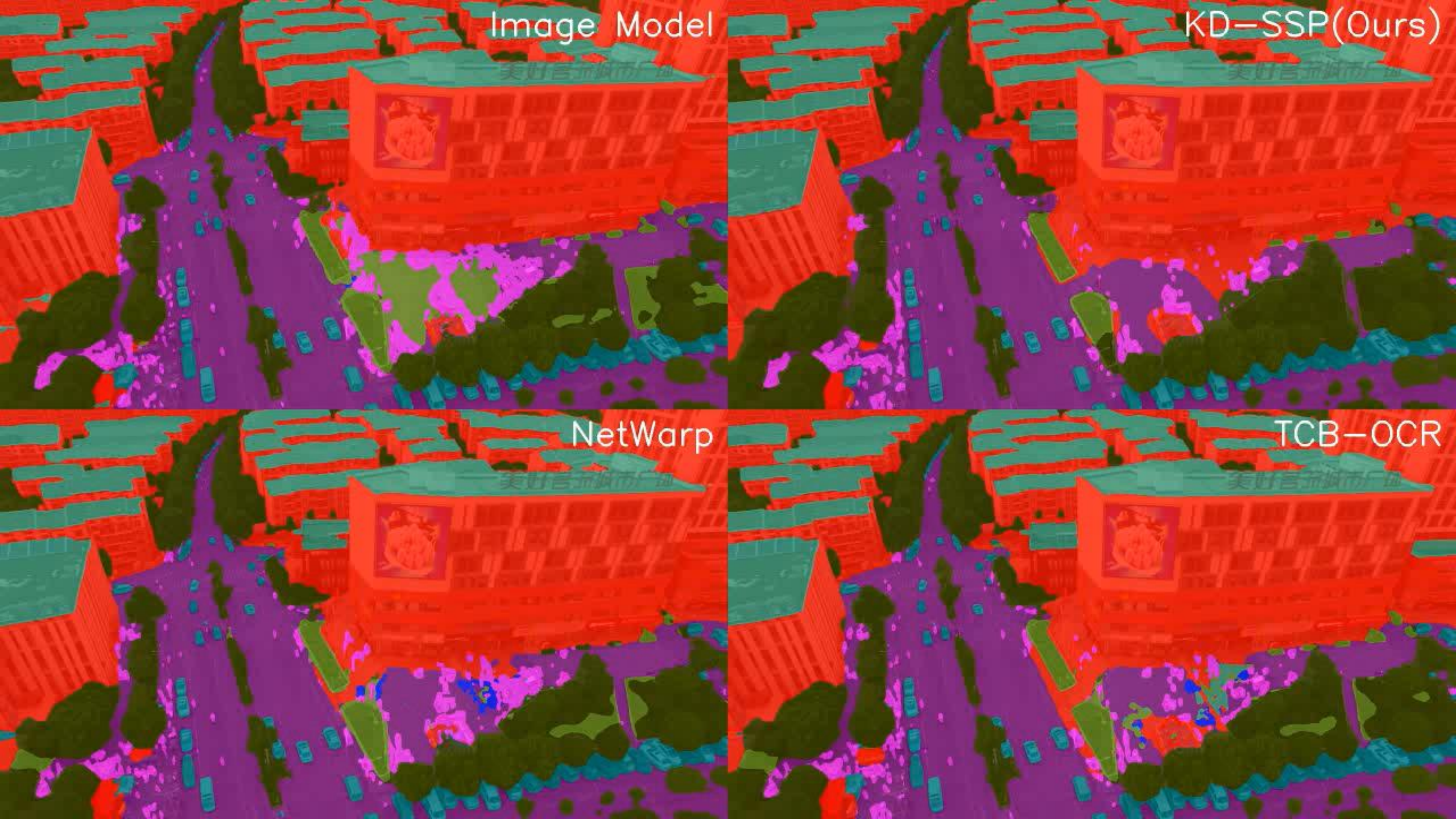


Image Model
KD-SSP(Ours)
NetWarp
TCB-OCR

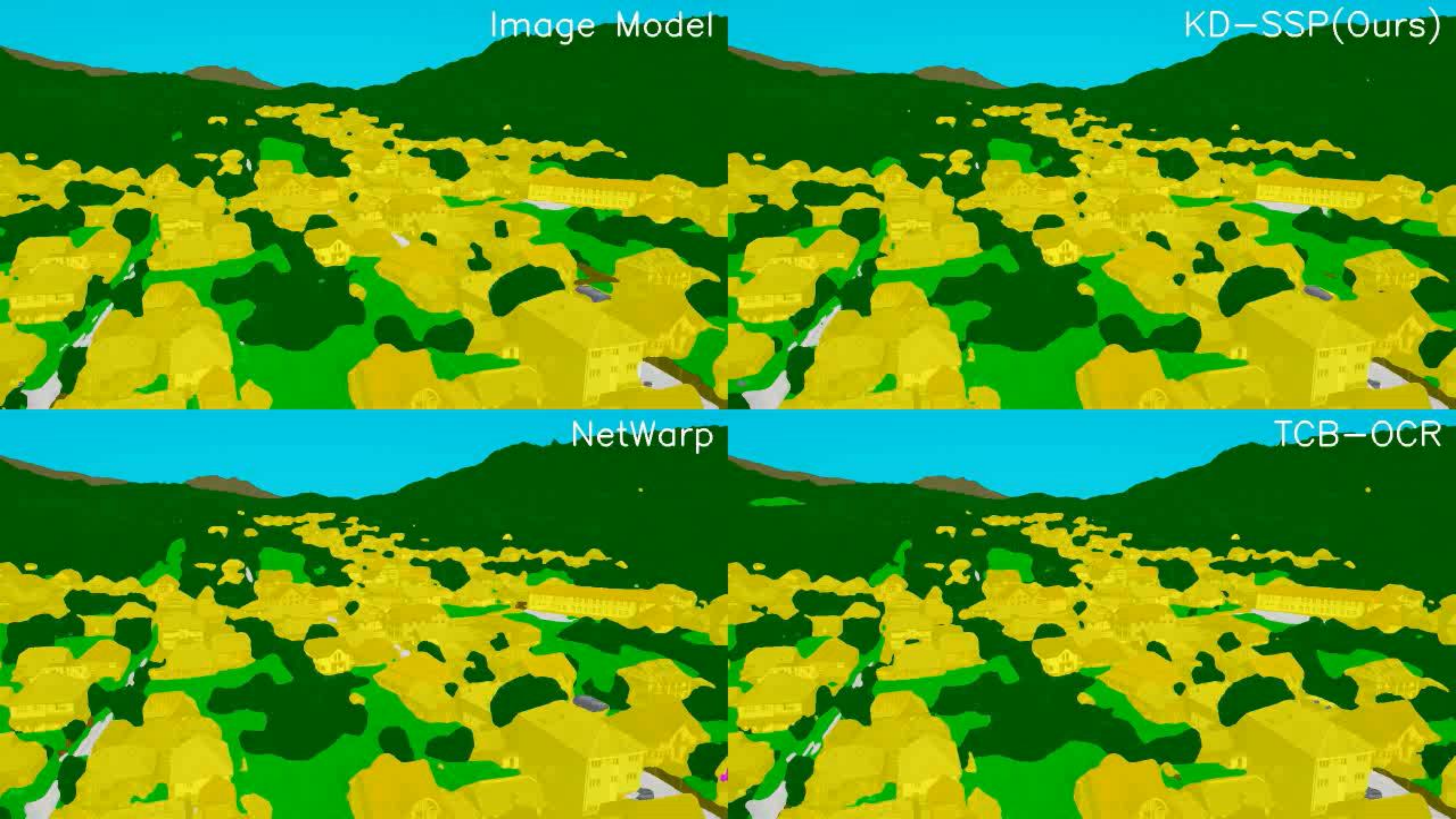


Image Model
KD-SSP(Ours)
NetWarp
TCB-OCR

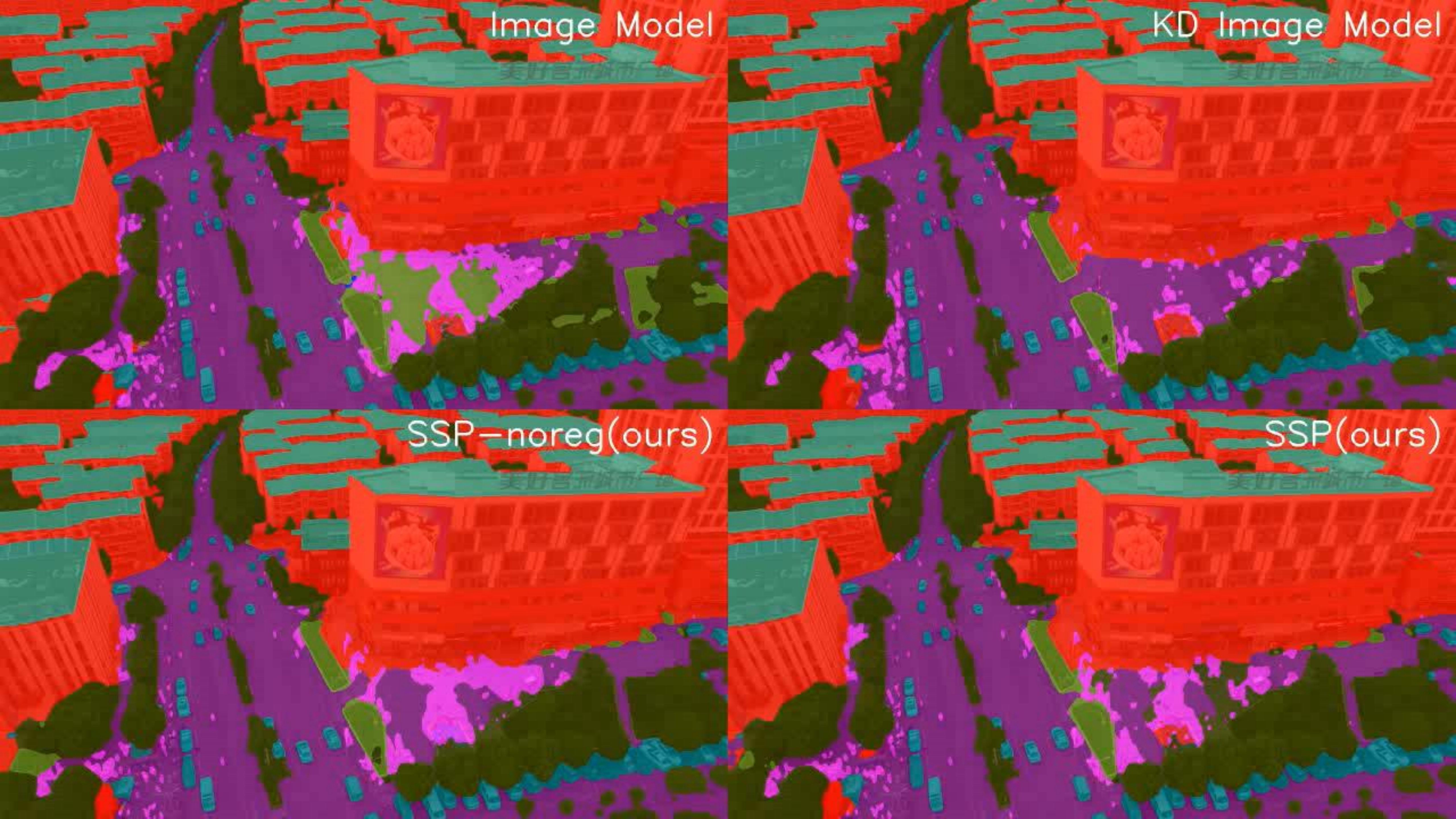


Image Model
KD Image Model
SSP-noreg(ours)
SSP(ours)

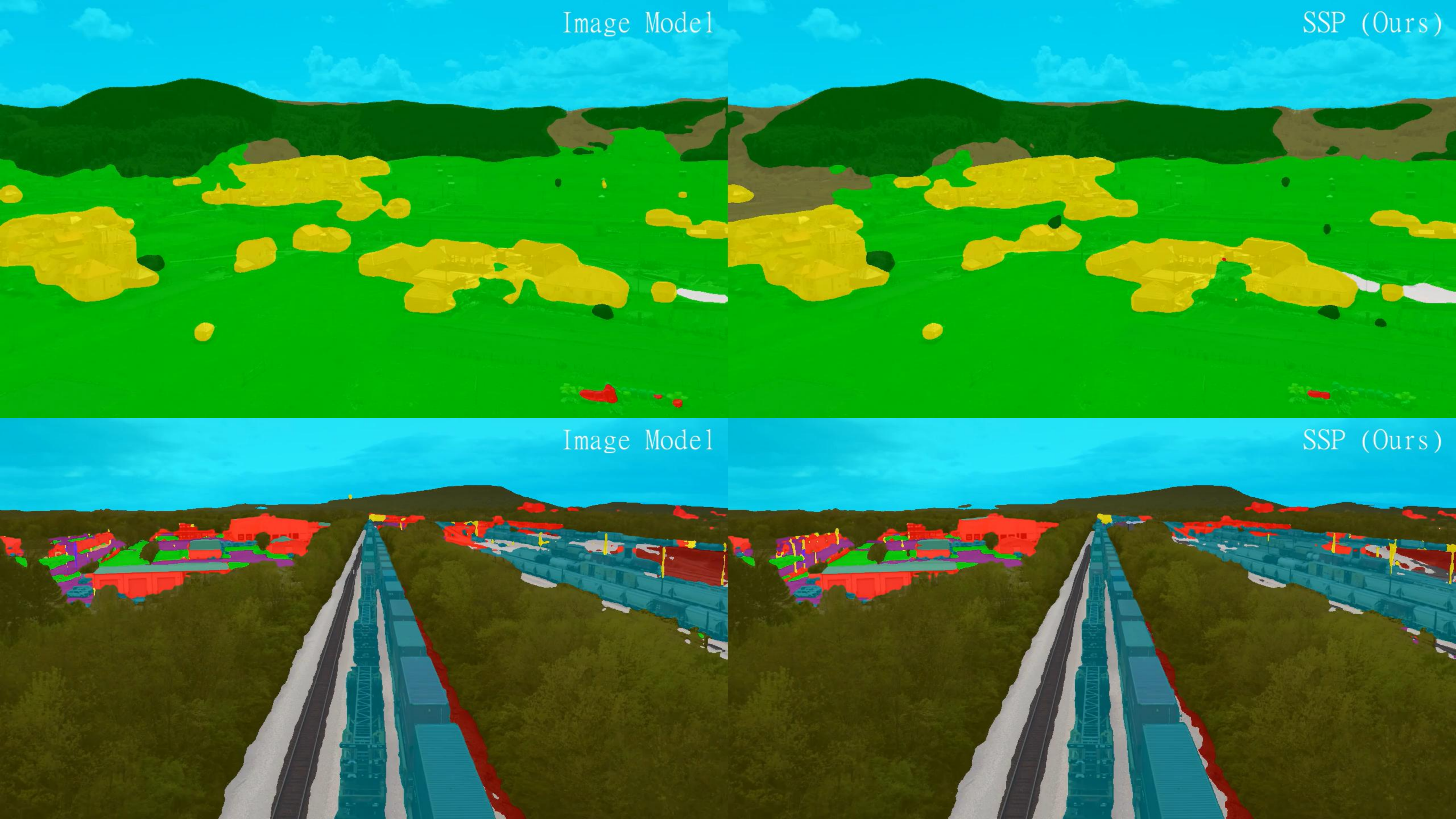


Image Model
SSP (Ours)
Image Model
SSP (Ours)

# **SSP**: High Temporal Consistency through **S**emantic **S**imilarity **P**ropagation in Semi-Supervised Video Semantic Segmentation for Autonomous Flight

- ✅ **SSP** propagates **semantic similarity** to ensure **temporal consistency**.
- ✅ **SSP** aligns predictions using **global transformations** to handle **UAV motion**.
- ✅ **KD-SSP** applies **consistency-aware distillation** to leverage **unlabeled frames** and boost **accuracy**.

**Project Page**

**IVI Homepage**

For additional **code** and **videos**, please visit our **project page:** https://gomtae.github.io/publications/ssp

For more **exciting projects**, please visit our **IVI homepage**: https://www.ivi.fraunhofer/en

**Cédric Vincent**

**Taehyoung Kim**

**Henri Meeß**