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
13vincentcedric@gmail.com, {taehyoung.kim, henri.meess}@ivi.fraunhofer.de

Abstract

*Semantic segmentation from RGB cameras is essential to the perception of autonomous flying vehicles. The stability of predictions through the captured videos is paramount to their reliability and, by extension, to the trustworthiness of the agents. In this paper, we propose a lightweight video semantic segmentation approach—suited to onboard real-time inference—achieving high temporal consistency on aerial data through **S**emantic **S**imilarity **P**ropagation across frames. **SSP** temporally propagates the predictions of an efficient image segmentation model with global registration alignment to compensate for camera movements. It combines the current estimation and the prior prediction with linear interpolation using weights computed from the features similarities of the two frames. Because data availability is a challenge in this domain, we propose a consistency-aware **K**nowledge **D**istillation training procedure for sparsely labeled datasets with few annotations. Using a large image segmentation model as a teacher to train the efficient SSP, we leverage the strong correlations between labeled and unlabeled frames in the same training videos to obtain high-quality supervision on all frames. **KD-SSP** obtains a significant temporal consistency increase over the base image segmentation model of* 12.5% *and* 6.7% *TC on UAVid and RuralScapes respectively, with higher accuracy and comparable inference speed. On these aerial datasets, KD-SSP provides a superior segmentation quality and inference speed trade-off than other video methods proposed for general applications and shows considerably higher consistency. Project page: https://github.com/FraunhoferIVI/SSP.*

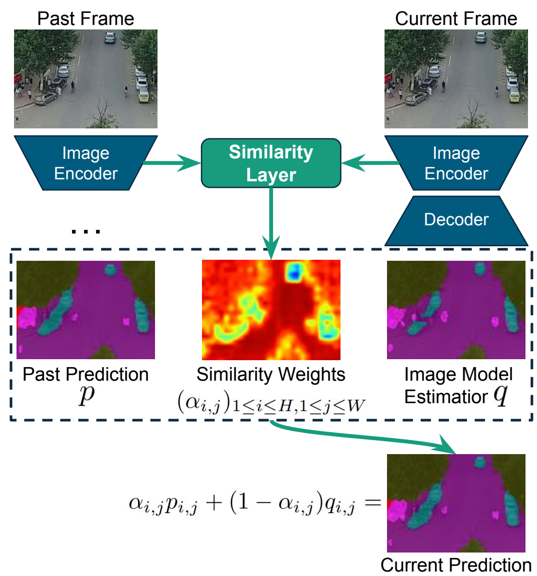

Figure 1. **Overview of Semantic Similarity Propagation (SSP)**. On each frame, the estimation of the image segmentation model is linearly combined with the last prediction based on interpolation weights computed from the feature similarities. This results in temporarily consistent robust predictions. The global registration alignment is omitted here.

1. Introduction

Autonomous Unmanned Aerial Vehicles (UAVs) rely on environmental perception for navigation. Semantic segmentation from RGB cameras serves as a crucial component for interpreting their surroundings, with additional use cases in emergency landing [16], 3D semantic reconstruction for scene understanding [49], or power lines detection [14]. It also has applications in other downstream tasks such as surveillance or photogrammetry [19, 21]. For trustworthy systems, the predictions must be temporally stable to avoid contradictory information or inconsistent predictions of crucial objects such as obstacles. While an image model can accurately process each frame independently, the temporal consistency of predictions requires a video model leveraging the correlations between consecutive frames.

Aerial footage captured at reasonable altitudes contains vastly different motion content compared to ground-level settings. With objects moving in a plane approximately or-

[*]Equal contribution

thogonal to the camera axis and at a large distance, most of the change between consecutive frames is a product of the camera motion. This results in a highly consistent semantic content requiring a specific temporal modeling method to be fully exploited. As an additional challenge, available data in this domain is scarce, and producing quality annotations of the complex scenes captured at high altitudes is resource-intensive. All publicly available datasets are sparsely labeled with few annotated frames per video: leveraging unlabeled frames in a semi-supervised setting is essential for better generalization.

Autonomous flight requires onboard, real-time processing of the video stream with high accuracy and consistency for reliability. However, most video segmentation approaches compromise on either inference speed or accuracy. Feature enhancement methods reach state-of-the-art accuracy through expensive temporal context modeling not suited to onboard applications [12, 33, 44]. Alternatively, key frame methods increase inference speed by reusing outputs from a subset of key frames to predict the intermediary non-key frames with fewer computations but inevitably less accuracy [41, 52, 55]. Both approaches are limited in prediction stability, as they do not encourage it in their training or architecture.

Our work aims to achieve reliable predictions—both accurate and stable—by leveraging temporal information across frames without sacrificing efficiency. For a lightweight approach suited to onboard applications, we explicitly **propagate** predictions to enforce consistency [26, 35, 36]. We linearly combine the output of an image segmentation model on the current frame with the previous frame's prediction at each time step, using per-pixel interpolation weights based on **semantic similarities** between the frames. Inspired by TFC [26], which uses cosine similarity on the feature maps to compute the weights, we extend their propagation method to entire videos using a convolutional similarity layer capable of learning the optimal balance between the past and current predictions without supervision from the labels. To further enhance consistency, we align the past prediction to the current frame using global registration to compensate for camera motion. This is significantly more efficient than optical flow alignment [35, 36] as the registration homography can be estimated from the UAV pose tracked by onboard sensors. While global registration assumes a planar ground, we show that our interpolation is robust to imperfect alignments in more complex real-world scenes. We complement our method with **knowledge distillation** for the semi-supervised training of our efficient model, utilizing a larger teacher model for supervision on all frames. ETC [28] introduced knowledge distillation for temporal consistency in an efficient image segmentation model, which we adapt to SSP. Instead of jointly training the teacher and student for consistency, we post-process the former's predictions with optical flow and restrain the losses to the prediction level because the teacher model cannot achieve the high consistency of explicitly propagated predictions.

In summary, we propose a new video semantic segmentation approach suitable for real-time onboard UAV applications without compromising accuracy, consistency, or efficiency. Our contributions are as follows:

- We predict linear interpolation weights with a convolutional similarity layer for a consistent and accurate propagation of predictions throughout videos, leveraging the semantic similarities of frames without supervision.
- We align previous predictions to the current frame using a global projective transformation, achieving stable combinations in aerial data dominated by camera motion and avoiding the high computational cost of optical flow.
- We train with a consistency-aware knowledge distillation, using unlabeled frames to prevent overfitting on the few annotated UAV frames and supervising the temporal propagation of predictions to enhance segmentation accuracy and consistency.

2. Related work

Image semantic segmentation. Since the introduction of FCN [42], encoder-decoder architectures with pre-trained backbones have been the standard for semantic segmentation. CNN models have evolved through larger spatial context modeling [5–7, 54], multi-scale representations [50], and relational context aggregation based on pixel correlations rather than spatial proximity [53]. Transformers further enhance this with self-attention [10, 47], and hierarchical transformers suited to dense predictions [29, 40, 51] now define state-of-the-art performance. Using image segmentation models on videos ignores the temporal correlations between frames, resulting in low consistency across predictions, especially for sparsely labeled datasets as few training samples lead to overfitting and a lack of regularization.

Video semantic segmentation. Real-time applications require online inference, which excludes methods leveraging bi-directional temporal interactions [1, 2, 31]. Most video methods are based on an encoder-decoder image segmentation model, introducing temporal interactions at the feature or prediction level. The two main approaches are accurate feature enhancement methods and efficient key frame methods. The former reaches higher accuracy by incorporating temporal context from adjacent frames into the static representations of the image encoder: STFCN [11] propagates features with a recurrent module, while NetWarp [12] uses optical flow and linear interpolation. Since TCB [33], most methods use attention-based mixing modules [18, 22, 23, 44, 48]. While these methods can scale beyond

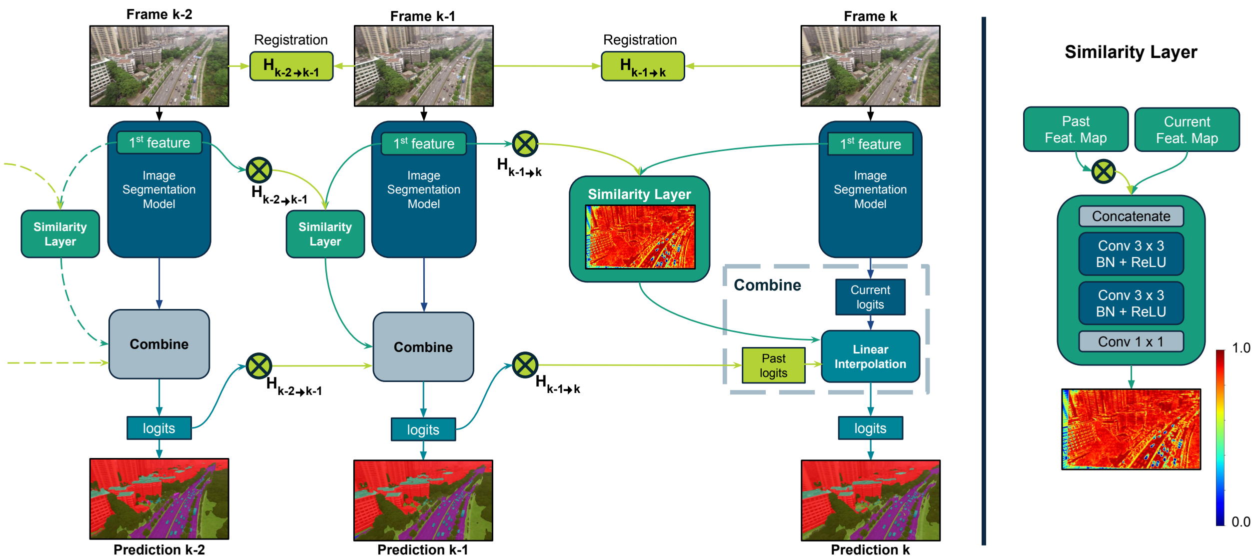

Figure 2. **Architecture of SSP, shown during inference over a video.** Predictions (as logits) are propagated through the whole video, with a global projective transformation H alignment to compensate for the camera movements. On each frame, the current estimation of the image semantic segmentation model is combined with the aligned past prediction by linear interpolation. The interpolation weights are computed from the extracted feature maps of the two frames by learnable convolutional layers. The first frame of each video is processed by the image model only. $\otimes$ represents the homography warping operation.

image models, they do not fit the computational and memory constraints of real-time on-edge inference. Key frame methods instead leverage temporal correlations to avoid redundant computations and achieve faster inference [41, 55]. Deep Feature Flow (DFF) [55] re-uses the feature maps of the last key frame aligned with optical flow. Variants include adaptive key frame selection [3, 24, 25, 52], methods for re-introducing non-key frame information [20, 24, 25], and optical flow alternatives with intrinsic learnable layers [24, 25]. Because these methods have inconsistent latency and unreliable non-key frame predictions, we focus on efficient image models and minimize the computational overhead of the temporal interactions used for consistency gains without compromising accuracy.

Temporal consistency of predictions. The stability of predictions can be explicitly enforced by combining the current frame's estimation from the base image segmentation model with those from previous frames. Miksik *et al.* [35] proposes a spatially local combination with the previous prediction after optical flow alignment. STGRU [36] further models consistency across frame sequences by propagating predictions with a gated recurrent unit and optical flow. However, the consistency gains of these methods are accompanied by a reduction in efficiency, mainly because of the slow optical flow computation. TFC [26] improves stability without compromising efficiency by using a simple linear interpolation between the current and past predictions without alignment, weighted by a similarity map computed with the cosine similarity of the low-level feature maps. As TFC is also a key frame method, consistency is only enforced between key frames, and predictions are not propagated through entire videos.

Contrastingly, ETC [28] processes each frame independently during inference with the image model and introduces temporal consistency during the training only through multiple consistency losses at both prediction and feature levels. Knowledge distillation from a teacher model is used to train efficient compact models. The resulting student model has improved temporal consistency through the regularization of representations over similar frames. As we enforce stability by propagating predictions, we use knowledge distillation to train on unlabeled frames instead.

Semi-supervised video semantic segmentation. Most video semantic segmentation datasets are sparsely labeled, as annotating every frame in a video is prohibitively resource-intensive. This is the case for all UAV datasets, which are additionally of smaller sizes. Semi-supervised methods tackle this challenge by leveraging unlabeled frames during training. Previously seen feature enhancement models train the image encoder on unlabeled adjacent frames to improve generalization and reduce overfitting on the few labeled frames [22, 57]. Orthogonally, pseudo-label methods generate supervision signals on unlabeled frames either in an offline manner [8, 13, 56] or online through self-regularization [43, 58] or the joint optimization of a teacher network with the same architecture [9]. Given that our efficient model focuses on the trade-off between prediction quality and inference speed instead of state-of-the-art accuracy, we propose a knowledge distillation training relying on a larger teacher model without any efficiency constraints. We leverage the temporal correlations between labeled and unlabeled frames in the training videos to provide high-quality supervision through inter-video overfitting. Feature enhancement and pseudo-label methods could then be used to train the teacher model for higher-quality supervision, although we do not experiment with them in this work.

3. Methodology

The architecture of SSP is illustrated in Fig. 2. We efficiently augment any image semantic segmentation model for stable inference across complete videos through a simple propagation of predictions with global registration alignment. Predictions on each frame are the linear combination of the image model estimation and the prior knowledge of the past frame's aligned prediction. The result is a more robust estimate for the current frame, more consistent with the past prediction. We propose a temporally consistent knowledge distillation procedure to train this efficient architecture on sparsely labeled datasets.

3.1. Temporal propagation of predictions

Predictions are propagated as logits. The image model output on the current frame is combined with the last frame prediction only as it is sufficient to enforce stability. The combination is a simple linear interpolation with weights representing the degree of semantic consistency between the two frames at each pixel's location.

Semantic similarity-based linear interpolation. With p the last prediction aligned to the current frame, and q the current output of the image model, the prediction of SSP is:

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

where $(\alpha_{i,j})_{1 \leq i \leq H, 1 \leq j \leq W}$ are the weights representing the semantic consistency on each pixel (i, j). These weights are computed by the similarity layer between the low-level feature maps of the past and current frame, illustrated in the right part of Fig. 2. We use convolutional layers on their concatenation for a local pattern-based detection of similarity between the semantic content of the two frames instead of the per-pixel cosine similarity of TFC [26]. Using more expressive learnable layers instead of a direct similarity measure facilitates the optimization of the predicted weights and minimizes the influence on the image encoder, which can focus on producing features suited to the segmentation task.

Global registration alignment. For the linear interpolation combination, the past prediction and feature map are aligned to the current frame with global registration to compensate for the camera movement only. We consider a homography H representing a projective transformation:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

where x and y are the coordinates in the original frame, and x' and y' are the coordinates in the aligned image. This transformation assumes a planar ground surface, which will result in alignment errors in more complex scenes. This does not result in a loss of accuracy as the similarity layer already expects misaligned pixels due to moving objects. Aligning the combined frames, even roughly, enables more pixels to corresponds and leads to higher consistency.

In practice, H is computed from the displacement (rotation and translation) of the UAV between the two frames, assuming that the ground is a flat plane and that the altitude is known [15, 34]. As this is unavailable in the considered datasets, we use a feature-based estimator matching keypoints between the frames [27, 46] for our experiments. This estimator and the differentiable warping operation are implemented by the Kornia library [38].

3.2. Video training

We first present a basic training procedure for SSP without knowledge distillation illustrated in Fig. 3, relying only on labeled samples. Training is done on two consecutive frames, propagating the prediction of the past frame $k - 1$ to the current one k on which the combination layers are trained. Only the current frame is assumed to be labeled, and the segmentation loss is computed on its prediction. To train for the temporal consistency of predictions in an unsupervised manner, as the past frame is not labeled, we use the same consistency loss as ETC [28]:

$$\mathcal{L}_{tc} = \frac{1}{HW} \sum_{i,j} \mathbf{O}_{i,j}^{k-1 \to k} \left\| y_{i,j} - \hat{x}_{i,j} \right\|_2^2 \tag{3}$$

where y is the current prediction, $\hat{x}$ the past prediction warped to the current frame with optical flow, and $\mathbf{O}^{k-1 \to k}$ an occlusion mask to reduce the influence of flow estimation errors and handle pixels which cannot be matched by optical flow: $\mathbf{O}_{i,j}^{k-1 \to k} = exp(- \left\| \mathbf{I}_{i,j}^{k} - \hat{\mathbf{I}}_{i,j}^{k-1} \right\|_1)$, with $\hat{\mathbf{I}}^{k-1}$ the past frame warped with optical flow to the current frame $\mathbf{I}^{k}$.

As the base image model is not frozen, it is trained on both frames from only the current frame's label. Similarly to ETC [28], this increases regularization through the consistency loss that ensures similar predictions on

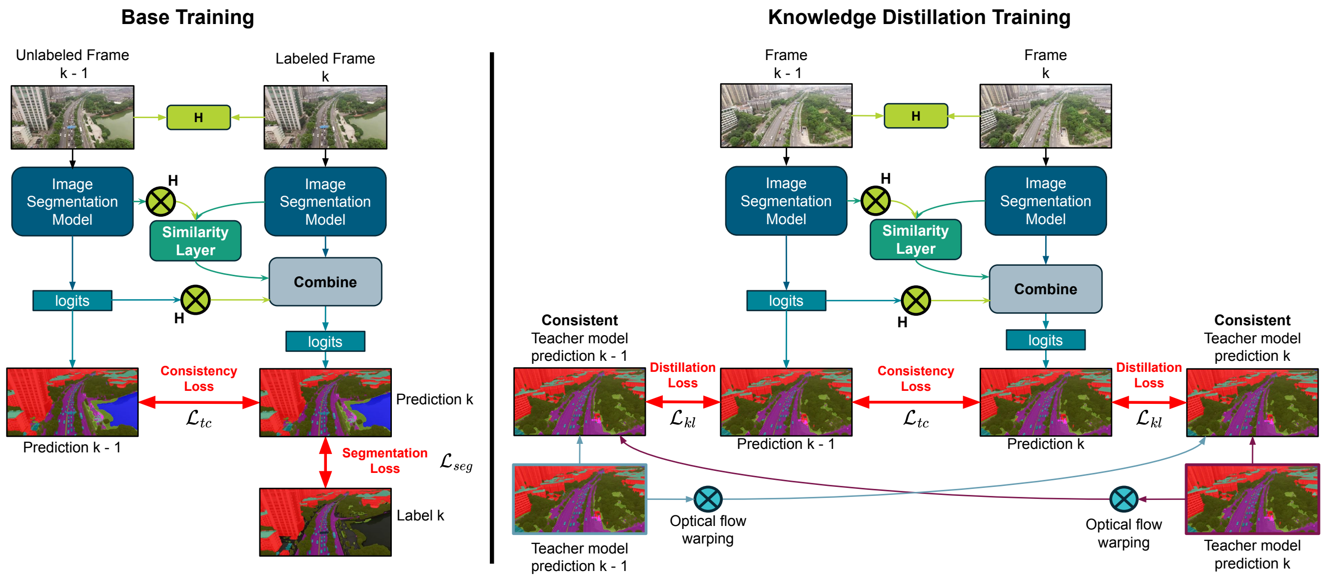

Figure 3. **SSP training process.** SSP is trained on two consecutive frames. The prediction of the past frame is propagated to the current frame on which the similarity layer is trained. The consistency loss is computed on the two predictions. Only the current frame needs to be labeled for the base training. With knowledge distillation, we train on all frames and add supervision on the past frame. An optical flow-based post-processing is used to make the teacher model's predictions consistent, as is expected of the student model's.

the two close frames. The total loss is $\mathcal{L}(P_q, P_p, A_q) = \mathcal{L}_{seg}(P_q, A_q) + \lambda\mathcal{L}_{tc}(P_q, P_p)$, where P_q and P_p are the current and past predictions respectively, and A_q the ground truth for the current frame. $\mathcal{L}_{seg}$ is the cross-entropy loss, and λ is the consistency loss weight.

3.3. Temporally consistent knowledge distillation

For semi-supervised training on sparsely labeled datasets, we propose to train on all frames through the supervision of a teacher model. This is a larger and more accurate image model trained on the annotated samples, which has no real-time inference efficiency constraints. We leverage the high similarity between frames of the same video to overfit this larger model to the training videos and produce high-quality soft annotations on all frames, unlabeled or not.

Our temporally consistent knowledge distillation training for SSP is illustrated in Fig. 3. We use the densely available teacher model predictions to supervise both frames with the distillation loss. However, as we also train for consistency, these predictions must be post-processed to be consistent. With T_q and T_p the teacher model's prediction on the current and past frames respectively, we use the optical flow between the two frames in both directions to obtain the consistent predictions used for supervision T_q^c and T_p^c:

$$T_p^c = (T_p + (1 - M_{occ})(W_{q\to p} * T_q))\frac{1}{2 - M_{occ}} \tag{4}$$

$$T_q^c = (T_q + (1 - M_{occ})(W_{p\to q} * T_p))\frac{1}{2 - M_{occ}} \tag{5}$$

where $W_{q\to p}$ and $W_{p\to q}$ are the optical flows between the frames in each direction, and $*$ is the optical flow warping operation. M_{occ} is a binary occlusion mask computed from a consistency check between the forward and backward flows [1, 39]: $M_{occ} = \|W_{q\to p} + W_{p\to q}\|_2^2 > 0.01(\|W_{q\to p}\|_2^2 + \|W_{p\to q}\|_2^2) + 0.5$.

For the segmentation loss, we use the knowledge distillation approach from [17], applying Kullback–Leibler divergence with a temperature parameter $\tau = 2$. The true labels are discarded, the teacher model is frozen, and we only supervise at the logits level: $\mathcal{L}_{kl}(P, T) = \text{softmax}(T/\tau)\log\left(\frac{\text{softmax}(T/\tau)}{\text{softmax}(P/\tau)}\right) \times \tau^2$, where P and T are the student and teacher logits, respectively. The distillation loss cannot be used at the feature level, as the consistency of the teacher model is not guaranteed. The loss function of the video knowledge distillation training is $\mathcal{L}(P_q, P_p, T_q^c, T_p^c) = \mathcal{L}_{kl}(P_q, T_q^c) + \mathcal{L}_{kl}(P_p, T_p^c) + \lambda_{kd}\mathcal{L}_{tc}(P_q, P_p)$.

4. Experiments

4.1. Datasets

We conduct our experiments on two UAV video semantic segmentation datasets. UAVid [30] is an urban dataset with 27 video sequences of 901 frames, captured at an altitude of around 50 meters and 20 frames per second (FPS) for 45 seconds each. It contains 270 annotated images, one every 100 frames or 5 seconds. Because each labeled sample requires a past frame to train SSP, we discard the first anno-

	Method	Params	GFLOPs	FPS A100	FPS Orin	UAVid mIoU ↑	UAVid TC ↑	RuralScapes mIoU ↑	RuralScapes TC ↑
	Teacher Model	**101.01M**	-	-	-	**81.92**	**84.09**	**66.65**	**89.43**
Image Models	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

Table 1. **Comparison on the validation sets of aerial datasets UAVid and RuralScapes.** Base Image Model is Hiera-S and UPerNet, Teacher Model is Hiera-B+ and UPerNet. Input resolution is 736 × 1280 except for the teacher model. While the feature enhancement methods outperform the base image model, SSP has the highest mIoU and TC. KD-SSP has a higher mIoU than the base image model trained with knowledge distillation (KD). Reported metrics on SSP and KD-SSP are the average of 3 runs. * For the optical flow computation in these models, we used the accurate but relatively slow RAFT [45], in the small configuration with 8 iterations. While a faster model would lead to higher inference speed, especially for DFF, it would result in even lower mIoU for DFF and an even more negligible improvement in mIoU and TC over the base image model for NetWarp.

tated frame of all videos and use only 243 annotated images. This corresponds to 180/63 labeled images and 18000/6300 total frames for training and validation, respectively. We use the annotations from [4], which include the addition of "roof" and "water" classes and the merging of "moving car" and "static car" for 8 semantic classes in total, to make this dataset more uniform with others for downstream applications. This re-mapping is only available for training and validation splits. Background is ignored during training. The RuralScapes dataset [32] is of comparable size to UAVid, with a focus on rural scenes. It is more densely annotated with one labeled frame per second for 1127 annotations. We sample this dataset at 10 FPS. As previously, we discard the first annotated frame of each video for training. This corresponds to 791/297 labeled images and 7139/3000 total frames for training and validation, respectively. RuralScapes contains 12 semantic classes different than those of urban scenes.

4.2. Evaluation metrics

Segmentation accuracy is measured with the mean Intersection over Union (mIoU) [42]. For stability, we require an unsupervised metric for sparsely labeled datasets and use the Temporal Consistency (TC) [28]. Given a pair of consecutive frames at time steps $k-1$ and k, it is defined as the mIoU between the first prediction P_{k-1} warped by optical flow $W_{k-1\to k}$ to the second frame, and the second prediction P_k: $TC_{k-1,k} = mIoU(W_{k-1\to k} * P_{k-1}, P_k)$, where $*$ is the optical flow warping operation. Then, the TC of a video is the mean of the TC of all consecutive pairs of frames. The inference speed is reported in frames per second (FPS) on an A100 (80G) and a Jetson Orin AGX (32G) GPU over 1000 samples after warmup, and only includes the forward pass. The global registration estimation is ignored for the speed benchmark of SSP, as it would be derived from the UAV pose in practice. We similarly report GFLOPs for an estimation of the computational cost.

4.3. Experimental details

Base image model. Any image semantic segmentation model producing pixel-wise classification logits can be used. We consider an encoder-decoder with Hiera-S [40] as the backbone, and UPerNet [50] with 256 channels for the segmentation head. We use the implementation and pretrained weights from the Segment Anything 2 (SAM2) encoder [37]. The results of the base image model serve as a baseline in our experiments. This model was chosen to obtain high segmentation accuracy efficiently on both datasets (see the comparison of image models in Tab. 1), allowing us to evaluate our video method's benefits on a precise model, as temporal consistency is intuitively harder to obtain on detailed predictions.

SSP. SSP is based on the base image model. We use the trained image models as starting points for our experiments to ensure easier and faster convergence, but this choice is challenged in the ablation study. SSP has 43.47M parameters, only 0.7% more than the base image model. For the similarity layer, we use the first extracted feature map of stride 4. Each convolutional layer divides the number of channels by 2. We set $\lambda = 0.5$ for the base training, and $\lambda_{kd} = 135000$.

Training parameters and teacher model. Except for the teacher model, the input resolution is set at 736×1280. The base image model is trained for 120 and 60 epochs on UAVid and RuralScapes respectively, with a batch size of 4, a learning rate of $2.5 \cdot 10^{-5}$, an AdamW optimizer with cosine scheduler and warmup, and 0.05 weight decay. For data augmentation, we use horizontal flipping, random scale and rotations, random noise, and random brightness/color changes. SSP is trained for 200 and 75 epochs on each dataset respectively, with double the learning rate and random square cropping to the shorter side in order to save memory. Other parameters are the same. For knowledge distillation, we use the same parameters and train for 20 and 25 epochs on UAVid and RuralScapes respectively.

The teacher model is trained for 200 and 75 epochs respectively, with the same parameters and random square cropping. For simplicity, we consider the same architecture in a larger size: Hiera-B+ and UPerNet with 512 channels, and with a larger input resolution of 1440×2560.

4.4. Results

We compare SSP to the base image model serving as a baseline and other video semantic segmentation methods [12, 33, 55] all using the same base encoder and decoder architecture. Results on both datasets are reported in Tab. 1. SSP with the base training yields 0.5% mIoU and 13% TC increases over the base image model on UAVid. On RuralScapes, results are similar with $+0.5\%$ mIoU and $+6.7\%$ TC. With knowledge distillation, KD-SSP obtains an additional $+0.8\%$ and $+0.56\%$ mIoU on UAVid and RuralScapes, respectively, for slightly lower TC than the base training.

We test the influence of knowledge distillation for semi-supervised learning by training the base image model with it. Results are referred to as the KD base image model. This corresponds to training on all frames from the distillation loss with the teacher model predictions. Temporal consistency is considerably improved, especially on UAVid which is more sparsely labeled, as training on a larger number of frames brings regularization for more uniform predictions on similar frames. Segmentation accuracy is also improved but still lower than KD-SSP, which also outperforms all other video models in mIoU and shows high temporal consistency due to the propagation of predictions.

SSP only has a marginally lower inference speed than the image model, resulting in a largely superior metrics/inference speed trade-off that is increased further for KD-SSP. Other video methods with worse metrics on these aerial datasets have considerably lower inference speeds due to expensive feature mixing for TCB and slow optical flow computation for DFF and NetWarp. The latter is because an accurate method of processing the frames at high resolution is needed for the small moving objects.

Model		mIoU	TC
Base image model		79.23	79.02
SSP	$\lambda = 0.2$	79.50	90.72
	$\lambda = 0.35$	**79.98**	91.58
	$\lambda = 0.5$	79.87	92.01
	$\lambda = 0.65$	79.47	**92.18**

Table 2. **Results with different weights of the consistency loss for our method.** Other experiments use $\lambda = 0.5$. UAVid dataset, 736×1280 resolution.

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.

4.5. Ablation study

We conduct an ablation study of our method on UAVid with the same parameters, omitting knowledge distillation to expedite the experimental process.

Consistency loss weight. The parameter λ balances segmentation and consistency losses. Results with different values of λ are in Tab. 2. Values between 0.35 and 0.5 appear optimal, as lower or higher λ result in worse metrics. This parameter does not control a simple trade-off between mIoU and TC as lower values lead to worse accuracy; however, it can be tuned to change the balance between the two as seen with $\lambda = 0.35$ and $\lambda = 0.5$.

Similarity layer. Using cosine similarity as TFC [26] instead of our proposed convolutions in the similarity layer, we obtain considerably inferior results and confirm that our choice is better suited to ensure long-term consistency through complete videos without loss of accuracy. Convolutional layers can learn to detect movements and change through local interactions, while the fixed per-pixel similarity measure is limited by the extracted feature maps.

Global registration alignment. Global registration is used to compensate for camera movements. Without it, SSP reaches the same mIoU and still significantly outperforms other methods in consistency. Less pixels are match-

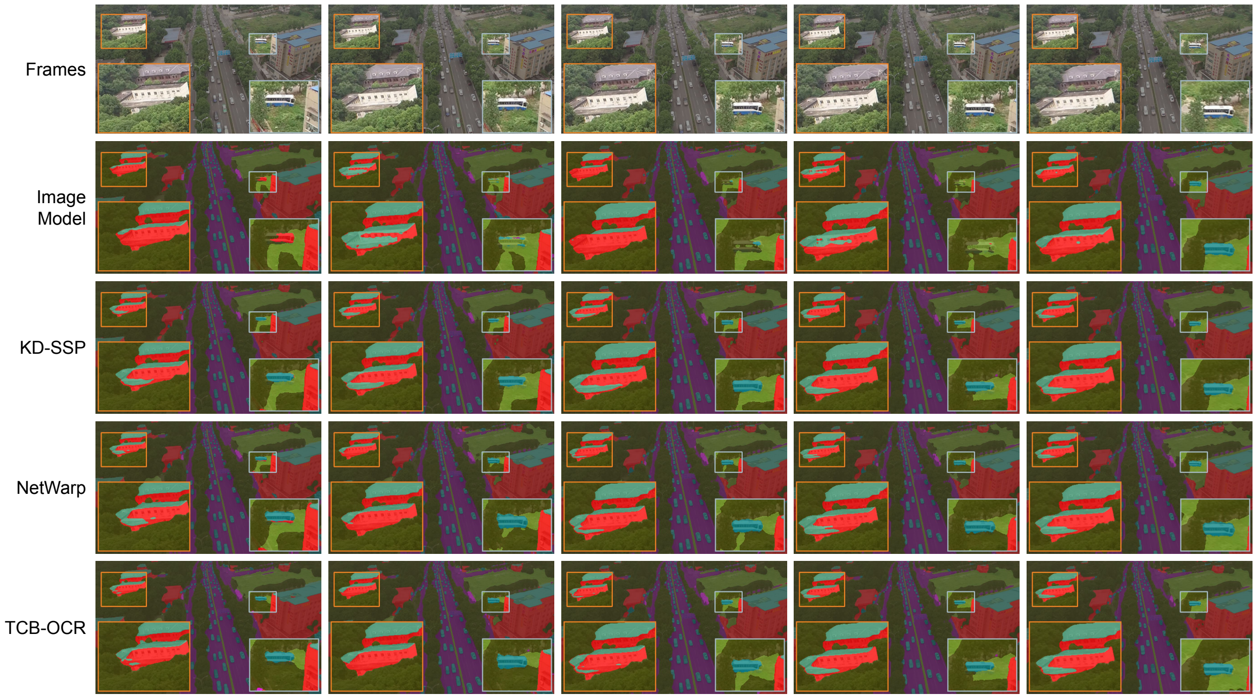

Figure 4. **Qualitative results.** Outputs of SSP compared to the base image model and other video methods [12, 33] on one video of the UAVid validation set, 736×1280 resolution. Frames are selected 5 frames apart. A comparison of results on entire videos showcasing the consistency gains are included in the supplementary material.

ing between the two frames without alignment, which explains the loss of TC. This experiment proves that SSP is robust to errors in the homography estimation as the similarity weights can adapt to misaligned frames and avoid loss of accuracy. SSP is still relevant when homography estimation from the UAV pose is unavailable and its computation from the frames is too slow, as it still outperforms all other methods in TC and mIoU.

Accuracy and consistency trade-off. We test the influence on accuracy of the two parts of our method enforcing consistency: the linear interpolation for the propagation of predictions and the consistency loss. Without interpolation, corresponding to training the base image model on pairs of frames with consistency loss, our method reaches comparable mIoU with an expected drop in TC. This proves that linear interpolation does not trade-off accuracy for consistency, and that training an image model with the consistency loss already presents high consistency gains as proposed by ETC [28]. Removing the consistency loss leads to the same conclusion, with even lower TC, showing that this loss is essential to learning temporally consistent predictions.

Base image model training. In all experiments, we used the trained base image model as the starting point for SSP, following a 2-step training: first, train the image model, then train SSP as proposed. This ensures easier and faster convergence without any compromise in the case of our experiments. However, in practice, training SSP in one step is more time-efficient, and we prove that this leads to the same metrics without additional training epochs.

5. Conclusion

This paper presents SSP, a new method for real-time aerial video semantic segmentation on-edge with high temporal consistency without loss of accuracy. SSP is suited to the motion content of drone-captured videos and to the computational restrictions of autonomous flight. The propagation of predictions with linear interpolation adds minimal computations on top of the base image segmentation model, for considerable gains in segmentation accuracy and especially temporal consistency. Global registration alignment is particularly suited to autonomous flight where it can be estimated from the UAV pose, and where compensating the camera movements leads to increased stability. If unavailable, SSP still outperforms existing image and video methods in accuracy, consistency, and metrics/inference speed trade-off. Alongside this model, we propose a temporally consistent knowledge distillation method for training SSP in a semi-supervised setting and confirm its relevance for efficient models by obtaining notable improvements in segmentation accuracy.

References

[1] Aharon Azulay, Tavi Halperin, Orestis Vantzos, Nadav Bornstein, and Ofir Bibi. Temporally stable video segmentation without video annotations. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1919–1928, 2022. 2, 5

[2] Razieh Kaviani Baghbaderani, Yuanxin Li, Shuangquan Wang, and Hairong Qi. Temporally-consistent video semantic segmentation with bidirectional occlusion-guided feature propagation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 674–684, 2024. 2

[3] Jiawen Cai, Yarong Liu, and Pan Qin. Attention based quick network with optical flow estimation for semantic segmentation. *IEEE Access*, 11:12402–12413, 2023. 3

[4] Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *arXiv preprint arXiv:2305.13608*, 2023. 6

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, page 833–851, Berlin, Heidelberg, 2018. Springer-Verlag. 2

[8] Liang-Chieh Chen, Raphael Lopes, Bowen Cheng, Maxwell Collins, Ekin Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. *Naive-Student: Leveraging Semi-Supervised Learning in Video Sequences for Urban Scene Segmentation*, pages 695–714. 2020. 3

[9] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2613–2622, 2021. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[11] Mohsen Fayyaz, Mohammad Hajizadeh Saffar, Mohammad Sabokrou, Mahmood Fathy, Fay Huang, and Reinhard Klette. Stfcn: Spatio-temporal fully convolutional neural network for semantic segmentation of street scenes. In *Computer Vision – ACCV 2016 Workshops*, pages 493–509, Cham, 2017. Springer International Publishing. 2

[12] Raghudeep Gadde, Varun Jampani, and Peter V. Gehler. Semantic video cnns through representation warping. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 6, 7, 8

[13] Aditya Ganeshan, Alexis Vallet, Yasunori Kudo, Shin-ichi Maeda, Tommi Kerola, Rares Ambrus, Dennis Park, and Adrien Gaidon. Warp-refine propagation: Semi-supervised auto-labeling via cycle-consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15499–15509, 2021. 3

[14] Gujing Han, Min Zhang, Qiang Li, Xia Liu, Tao Li, Liu Zhao, Kaipei Liu, and Liang Qin. A lightweight aerial power line segmentation algorithm based on attention mechanism. *Machines*, 10(10), 2022. 1

[15] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004. 4

[16] Philip Hausmann, Henri Meeß, and Gordon Elger. Image segmentation based emergency landing for autonomous and automated unmanned aerial vehicles. In *33th Congress of the International Council of the Aeronautical Sciences*, 2022. 1

[17] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 5

[18] Yuanduo Hong, Huihui Pan, Weichao Sun, Yulei Wang, Ning Bian, and Huijun Gao. Dual spatial-temporal feature pyramid with decoupled temporal mining for video semantic segmentation. *IEEE Transactions on Intelligent Vehicles*, pages 1–12, 2024. 2

[19] SM Islam. Drones on the rise: Exploring the current and future potential of uavs. *arXiv preprint arXiv:2304.13702*, 2023. 1

[20] Samvit Jain, Xin Wang, and Joseph E. Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8858–8867, 2019. 3

[21] San Jiang, Wanshou Jiang, and Lizhe Wang. Unmanned aerial vehicle-based photogrammetric 3d mapping: A survey of techniques, applications, and challenges. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):135–171, 2022. 1

[22] Jiangwei Lao, Weixiang Hong, Xin Guo, Yingying Zhang, Jian Wang, Jingdong Chen, and Wei Chu. Simultaneously short- and long-term temporal modeling for semi-supervised video semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14763–14772, 2023. 2, 3

[23] Jiangtong Li, Wentao Wang, Junjie Chen, Li Niu, Jianlou Si, Chen Qian, and Liqing Zhang. Video semantic segmentation via sparse temporal transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 59–68, New York, NY, USA, 2021. Association for Computing Machinery. 2

[24] Yule Li, Jianping Shi, and Dahua Lin. Low-latency video semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5997–6005, 2018. 3

[25] Zhixue Liang, Wenyong Dong, and Bo Zhang. A dual-branch hybrid network of cnn and transformer with adaptive keyframe scheduling for video semantic segmentation. *Multimedia Systems*, 30, 2024. 3

[26] Matthieu Lin, Jenny Sheng, Yubin Hu, Yangguang Li, Lu Qi, Andrew Zhao, Gao Huang, and Yong-Jin Liu. Exploring temporal feature correlation for efficient and stable video semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3450–3458, 2024. 2, 3, 4, 7

[27] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 4

[28] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *Computer Vision – ECCV 2020*, pages 352–368, Cham, 2020. Springer International Publishing. 2, 3, 4, 6, 8

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[30] 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. 5

[31] Behrooz Mahasseni, Sinisa Todorovic, and Alan Fern. Budget-aware deep semantic video segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2077–2086, 2017. 2

[32] 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. 6

[33] Jiaxu Miao, Yunchao Wei, Yu Wu, 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. 2, 6, 7, 8

[34] Michael Miksch, Bin Yang, and Klaus Zimmermann. Automatic extrinsic camera self-calibration based on homography and epipolar geometry. In *2010 IEEE Intelligent Vehicles Symposium*, pages 832–839, 2010. 4

[35] Ondrej Miksik, Daniel Munoz, J. Andrew Bagnell, and Martial Hebert. Efficient temporal consistency for streaming video scene analysis. In *2013 IEEE International Conference on Robotics and Automation*, pages 133–139, 2013. 2, 3

[36] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6819–6828, 2018. 2, 3

[37] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6

[38] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch, 2019. 4

[39] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. *Artistic Style Transfer for Videos*, page 26–36. Springer International Publishing, 2016. 5

[40] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. *ICML*, 2023. 2, 6

[41] Evan Shelhamer, Kate Rakelly, Judy Hoffman, and Trevor Darrell. Clockwork convnets for video semantic segmentation. In *Computer Vision – ECCV 2016 Workshops*, pages 852–868, Cham, 2016. Springer International Publishing. 2, 3

[42] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017. 2, 6

[43] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, pages 596–608. Curran Associates, Inc., 2020. 3

[44] Guolei Sun, Yun Liu, Henghui Ding, Min Wu, and Luc Van Gool. Learning local and global temporal contexts for video semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6919–6934, 2024. 2

[45] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, pages 402–419, Cham, 2020. Springer International Publishing. 6

[46] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33, 2020. 4

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 2

[48] Hao Wang, Weining Wang, and Jing Liu. Temporal memory attention for video semantic segmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2254–2258. IEEE, 2021. 2

[49] Zizhuang Wei, Yao Wang, Hongwei Yi, Yisong Chen, and Guoping Wang. Semantic 3d reconstruction with learning mvs and 2d segmentation of aerial images. *Applied Sciences*, 10(4), 2020. 1

[50] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Computer Vision – ECCV 2018*, pages 432–448, Cham, 2018. Springer International Publishing. 2, 6

[51] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, pages 12077–12090. Curran Associates, Inc., 2021. 2

[52] Yu-Shuan Xu, Hsuan-Kung Yang, Tsu-Jui Fu, and Chun-Yi Lee. Dynamic video segmentation network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[53] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision – ECCV 2020*, pages 173–190, Cham, 2020. Springer International Publishing. 2

[54] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2

[55] 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. 2, 3, 6, 7

[56] Yi Zhu, Karan Sapra, Fitsum A. Reda, Kevin J. Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8848–8857, 2019. 3

[57] Jiafan Zhuang, Zilei Wang, and Yuan Gao. Semi-supervised video semantic segmentation with inter-frame feature reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3263–3271, 2022. 3

[58] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. In *International Conference on Learning Representations*, 2021. 3