35th British Machine Vision Conference (BMVC), Glasgow, UK, 2024.

MCDS-VSS: Moving Camera Dynamic Scene Video Semantic Segmentation by Filtering with Self-Supervised Geometry and Motion

Angel Villar-Corrales
villar@ais.uni-bonn.de
Moritz Austermann
austermann@ais.uni-bonn.de
Sven Behnke
behnke@cs.uni-bonn.de

Autonomous Intelligent Systems –
Computer Science Institute VI,
Center for Robotics and
Lamarr Institute for Machine Learning
and Artificial Intelligence,
University of Bonn, Germany

Abstract

Autonomous systems, such as self-driving cars, rely on reliable semantic environment perception for decision making. Despite great advances in video semantic segmentation, existing approaches ignore important inductive biases and lack structured and interpretable internal representations. In this work, we propose *MCDS-VSS*, a structured filter model that learns in a self-supervised manner to estimate scene geometry and ego-motion of the camera, while also estimating the motion of external objects. Our model leverages these representations to improve the temporal consistency of semantic segmentation without sacrificing segmentation accuracy. MCDS-VSS follows a prediction-fusion approach in which scene geometry and camera motion are first used to compensate for ego-motion, then residual flow is used to compensate motion of dynamic objects, and finally the predicted scene features are fused with the current features to obtain a temporally consistent scene segmentation. Our model parses automotive scenes into multiple decoupled interpretable representations such as scene geometry, ego-motion, and object motion. Quantitative evaluation shows that MCDS-VSS achieves superior temporal consistency on video sequences while retaining competitive segmentation performance. Code and pretrained models are available in the project website.

1 Introduction

Video semantic segmentation (VSS) is the task of assigning a categorical label to each pixel in every frame of a video sequence [56]. This task is highly relevant in the field of robotics, where understanding and interpreting scenes from video is crucial for many applications, e.g. autonomous driving [39] or indoor service tasks [35]. Thanks to the availability of high-quality image datasets, semantic segmentation of automotive scenarios has recently seen tremendous progress [6, 48, 55]. However, obtaining temporally consistent segmentation of video sequences still remains a challenge due to the lack of large-scale annotated video datasets and the lack of suitable inductive biases for video processing.

© 2024. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

To address these limitations, existing VSS models enforce temporal continuity by propagating features across multiple frames through the use of unstructured recurrent networks [32, 38], optical flow models [9, 10], or transformers [25, 47]; thus exploiting temporal correlations in the video sequences in a data-driven manner.

However, these models ignore specific properties from the target domain, which could potentially be incorporated into the model architecture in order to improve its performance and generalization capabilities. For instance, in the automotive domain, the observations taken from a moving vehicle can be decomposed into *static background features*, which move only due to the ego-motion of the vehicle, and *dynamic object features* that correspond to moving objects. Incorporating such motion and geometric inductive biases into the network architecture can lead to models producing a more temporally consistent interpretation of the scene, outperforming models that attempt to learn these properties solely from data.

To test this hypothesis, we propose *MCDS-VSS*, a structured recurrent model that explicitly incorporates geometry and motion inductive biases from the *moving camera dynamic scene* (MCDS) domain in order to improve the temporal consistency of a segmentation network. MCDS-VSS follows a prediction-fusion approach in which ego-motion is compensated by projecting scene features into the current time-step using estimated scene geometry and estimated camera motion. Estimated residual flow is then used to compensate for object motion. Finally, the predicted features are fused with the features extracted from the current frame to obtain a temporally consistent semantic segmentation of the scene.

Through self-supervised learning (SSL), MCDS-VSS learns to estimate scene geometry and ego-motion. It also estimates motion of additional moving objects (e.g. pedestrians or vehicles), and hard-wires our knowledge from the MCDS domain to project the previous scene features into the current time-step using these representations. The structured design of our filter allows us to factorize the perceived complex changes in the scene into simpler factors of variation; thus easing the modeling of temporal information.

Our experiments show that MCDS-VSS improves the temporal consistency of a segmentation model without compromising its segmentation performance, outperforming VSS baselines which ignore moving camera dynamic scene inductive biases, and performing comparably to state-of-the-art VSS models. Furthermore, MCDS-VSS parses an automotive scene into interpretable internal representations, such as depth, camera motion, and object flow.

In summary, our contributions are as follows:

- We propose MCDS-VSS, a structured recurrent filter that improves the temporal consistency of a segmentation model without sacrificing segmentation performance.
- MCDS-VSS learns depth and ego-motion in a self-supervised way, and uses these representations together with estimated object motion to propagate scene features.
- Our model outperforms existing VSS baselines on Cityscapes—achieving superior temporal consistency and parsing the scene into human-interpretable representations.

2 Related Work

Video Semantic Segmentation: VSS methods are often divided into two distinct categories. The first class aims to reduce the computational cost and improve the efficiency of segmentation models, instead of naively encoding and interpreting every single input frame. Several methods improve the efficiency by propagating and reusing features extracted from selected key frames [20, 59]; whereas other approaches achieve efficiency by employing lightweight neural network blocks [31, 36] or by distilling the information from large teacher models into smaller models [27]. The second category, to which our proposed method belongs, aims to improve the semantic segmentation performance and temporal consistency by exploiting the temporal continuity of video streams. Some methods exploit temporal dependencies between video frames and improve the consistency of the predicted segmentation maps by combining image segmentation models with recurrent neural networks (RNNs) [32, 36, 38, 44] or with attention-based modules [23, 25, 40, 41, 47]. Another family of works use an optical flow module to compute the feature correspondence between consecutive frames, and then use this flow for predictive feature learning [2, 9, 19, 27, 43, 50, 59].

Our method belongs to the latter category of VSS models. However, unlike aforementioned approaches, MCDS-VSS incorporates assumptions from the domain of moving cameras and dynamic scenes into the model design, and computes interpretable intermediate geometry and motion-aware representations, which lead to accurate and temporally consistent video segmentation results.

Improving Segmentation via Depth & Camera Motion Estimation: Self-supervised depth estimation (SSDE) aims to learn the scene geometry from unlabeled monocular videos, without any recorded depth information. This is often achieved by training a neural network to jointly predict the scene depth and camera ego-motion between two video frames, synthesizing the second frame from the first using differentiable warping, and minimizing a photometric loss function [4, 11, 14, 46, 57].

The interplay between semantic segmentation and SSDE has been studied for various tasks, including depth estimation [52], domain adaptation [15, 22, 49], and semi-supervised learning [17, 29]. These models exploit SSDE as an additional source of supervision, helping segmentation models learn high-level semantic features, especially when few labeled samples are available.

Several works have investigated the use of depth and motion for semantic segmentation in videos. Approaches like [1, 5, 37, 58] segment dynamic scenes by jointly processing video frames with depth information captured by LiDAR scanners; whereas methods like [7, 21, 28, 34] use depth and camera pose information in combination with a semantic segmentation model in order to improve the segmentation performance by enforcing consistency between predictions from multiple viewpoints. Recently, depth-aware panoptic segmentation models [30, 33, 51] aim to jointly solve the tasks of panoptic segmentation and depth estimation by extending a segmentation model with a depth decoder and conditioning its prediction using instance-masks.

The method most similar to ours is Wagner *et al.* [45], which leverages depth and camera motion learned in a supervised manner to improve the performance of a segmentation model on video sequences. However, this method has several limitations, including not modeling moving objects and requiring ground truth depth and poses, thus limiting its applicability. In contrast, MCDS-VSS addresses the limitations, being able to process challenging dynamic scenes with moving cameras, even in the absence of depth information and camera poses.

3 MCDS-VSS Structured Filtering Method

We propose MCDS-VSS, illustrated in Figure 1, a structured filter that improves the temporal consistency of a semantic segmentation model on moving camera dynamic scene scenarios.

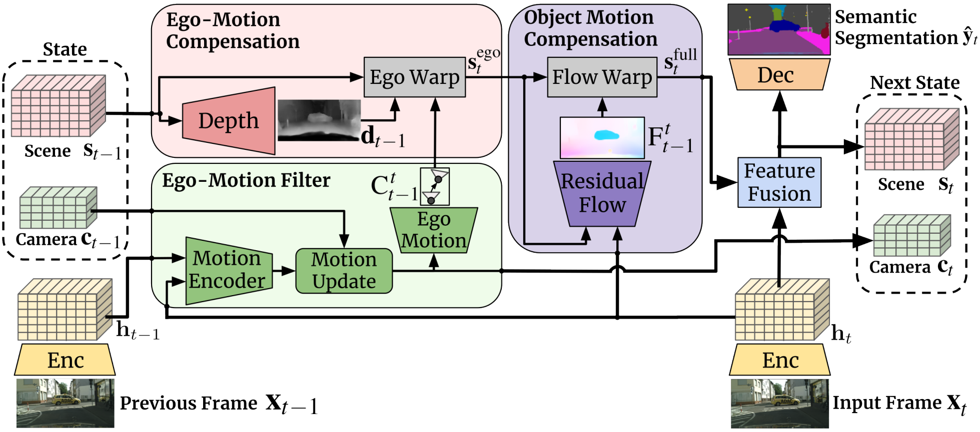


Figure 1: MCDS-VSS structured filter. Scene depth $\mathbf{d}_{t-1}$, ego-motion C_{t-1}^t, and object-motion F_{t-1}^t are used to project scene features $\mathbf{s}_{t-1}$ to the current time t, where they are fused with current image features $\mathbf{h}_t$ to obtain a temporally consistent semantic segmentation $\hat{\mathbf{y}}_t$.

MCDS-VSS learns in a self-supervised way to estimate geometry and motion representations, i.e., scene depth and camera ego-motion. It also estimates the motion of other agents in the scene, and uses these human-interpretable representations to propagate abstract scene features over time, thus improving its segmentation performance and temporal consistency.

MCDS-VSS is composed of an image encoder $\mathcal{E}_{\text{x}}$, a structured filter, and a segmentation decoder $\mathcal{D}_{\text{y}}$. It receives as input a sequence of RGB frames $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$ and encodes them into feature maps $\{\mathbf{h}_1, ..., \mathbf{h}_T\}$, which are then recursively processed to integrate temporal information, and decoded into semantic segmentation maps $\mathcal{Y} = \{\hat{\mathbf{y}}_1 ..., \hat{\mathbf{y}}_T\}$.

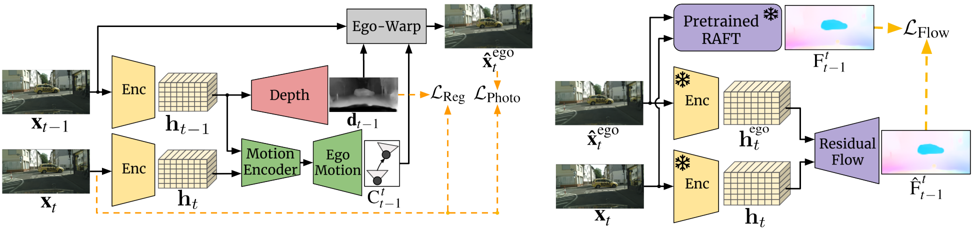
3.1 Learning of Geometry & Motion

MCDS-VSS learns in a self-supervised manner to estimate scene geometry and camera motion, which are then used to improve the temporal consistency of a segmentation model. Figure 2 illustrates our two-step self-supervised approach for learning the scene geometry with camera motion and for distillation of object dynamics.

Scene Geometry and Ego-Motion: We train our model to predict the monocular depth and camera pose transformation of the vehicle in a self-supervised manner by solving a novel view-synthesis pretext task in which a target image $\mathbf{x}_t$ is rendered from a source image $\mathbf{x}_{t-1}$ by modeling the static scene features that change due to the ego-motion [13, 57].

To predict the scene geometry, MCDS-VSS incorporates a *depth decoder* $\mathcal{D}_{\text{d}}$ that outputs the depth $\mathbf{d}_t$ and inverse depth $\mathbf{d}_t^{\triangleleft}$[1] of the scene given the input feature maps $\mathbf{h}_t$; whereas to compute the camera motion between two images we employ a *motion encoder* $\mathcal{E}_{\text{m}}$ that computes motion features between two sets of feature maps, and an *ego-motion decoder* $\mathcal{D}_{\text{c}}$, which predicts the camera transformation between two time steps C_{t-1}^t, parameterized as a 6-dimensional vector containing the translation parameters and Euler angles of the camera transformation matrix. We then render the ego-warped image $\hat{\mathbf{x}}_t^{\text{ego}}$ using the estimated scene

[1]With slight abuse of notation, we denote the $\mathbf{d}_t^{\triangleleft}$ as *disparities*.



(a) SSL of depth and ego-motion. (b) Distillation of object-motion.

Figure 2: Learning geometry and motion. **a)** We learn the scene depth $\mathbf{d}_{t-1}$ and ego-motion C_{t-1}^{t} in a self-supervised manner given two video frames by enforcing a photometric loss $\mathcal{L}_{\text{Photo}}$ between the ego-warped $\hat{\mathbf{x}}_t^{\text{ego}}$ and target frames $\mathbf{x}_t$, as well as a depth regularization $\mathcal{L}_{\text{Reg}}$. **b)** Given an ego-warped image, we train a residual flow decoder to predict the residual optical flow $\hat{\mathrm{F}}_{t-1}^{t}$ that parameterizes the dynamics of moving objects in the scene by distilling a pretrained RAFT model.

depth and ego-motion:

$$\hat{\mathbf{x}}_t^{\text{ego}} = \mathcal{F}_{\text{fwd}}(\mathbf{x}_{t-1}, \mathbf{d}_{t-1}, \mathrm{C}_{t-1}^{t}, K), \tag{1}$$

where $K \in \mathbb{R}^{3,3}$ are the camera intrinsic parameters, and $\mathcal{F}_{\text{fwd}}$ is the forward rendering function proposed in [24]. If the depth and ego-motion estimates are accurate, the resulting warped images should match the target images except for occluded regions and moving objects. Therefore, as illustrated in Figure 2a, we train the modules for self-supervised depth and ego-motion estimation by optimizing the following loss function:

$$\mathcal{L}_{\text{Depth}} = \mathcal{L}_{\text{Photo}}(\hat{\mathbf{x}}_t^{\text{ego}}, \mathbf{x}_t) + \lambda_{\text{Reg}} \cdot \mathcal{L}_{\text{Reg}}(\mathbf{d}_{t-1}^{\triangleleft}), \tag{2}$$

$$\mathcal{L}_{\text{Photo}} = \frac{\alpha}{2}(1 - \text{SSIM}(\hat{\mathbf{x}}_t^{\text{ego}}, \mathbf{x}_t)) + (1-\alpha)||\hat{\mathbf{x}}_t^{\text{ego}} - \mathbf{x}_t||_1, \tag{3}$$

$$\mathcal{L}_{\text{Reg}} = |\partial_x \widetilde{\mathbf{d}}_t^{\triangleleft}| e^{-|\partial_x \mathbf{x}_t|} + |\partial_y \widetilde{\mathbf{d}}_t^{\triangleleft}| e^{-|\partial_y \mathbf{x}_t|}, \tag{4}$$

where ∂_x and ∂_y are the spatial gradients in the x- and y-directions, SSIM is the structural similarity index, and $\widetilde{\mathbf{d}}^{\triangleleft}$ is the normalized disparity map. $\mathcal{L}_{\text{Photo}}$ is a photometric loss that measures the difference between the ego-warped and target images, and $\mathcal{L}_{\text{Reg}}$ is an edge-aware smoothing regularization [12] that encourages the normalized disparity maps to be locally smooth, except on the image edges. To mitigate the effect of disocclusions and moving objects during training, we use the auto-masking and per-pixel minimum processing steps proposed in [13].

Object Motion: Assuming static scenes as well as accurate depth and ego-motion estimates, the predicted ego-warped images $\hat{\mathbf{x}}_t^{\text{ego}}$ are identical, up to occluded regions, to the target images $\mathbf{x}_t$. Hence, we make the assumption that any major differences between such frames must be explained by external moving objects (e.g. driving cars or pedestrians).

As illustrated in Figure 2b, we estimate the residual optical flow $\hat{\mathrm{F}}_{t-1}^{t}$ between the ego-warped and target images, which encodes the dynamics of moving objects, by training a *residual flow decoder* $\mathcal{R}_{\text{f}}$ while keeping all other modules frozen. The residual flow $\hat{\mathrm{F}}_{t-1}^{t}$ is parameterized as a 2D flow field that encodes the per-pixel motion in the horizontal and vertical directions needed to align the ego-warped images $\hat{\mathbf{x}}_t^{\text{ego}}$ to the corresponding target images $\mathbf{x}_t$. $\mathcal{R}_\mathrm{f}$ is trained to match the optical flow predictions of the large state-of-the-art optical flow model RAFT [42]:

$$\mathcal{L}_\mathrm{Flow} = ||\hat{\mathrm{F}}_{t-1}^{t} - \mathrm{F}_{t-1}^{t}||_1. \tag{5}$$

3.2 Structured Filter

The modules and representations described in Section 3.1 form the core of the MCDS-VSS structured filter, which is depicted in Figure 1. It propagates information over time using two different filter states, namely a *scene state* $\mathbf{s}$ that encodes the scene contents and geometry, and a *camera state* $\mathbf{c}$ that encodes the ego-motion of the vehicle.

It consists of six components: ego-motion filter, depth estimation, ego-motion compensation, residual flow estimation, object motion compensation, and feature fusion.

Ego-Motion Filter: The ego-motion filter extends the motion encoder $\mathcal{E}_\mathrm{m}$ and ego-motion decoder $\mathcal{D}_\mathrm{c}$ modules in order to aggregate motion information over time and enforce the prediction of temporally consistent ego-motion. The temporal integration is achieved via a *motion update* module, which is implemented as a ConvGRU [3] recurrent layer that jointly processes the motion features and previous camera state $\mathbf{c}_{t-1}$, and outputs an updated camera state $\mathbf{c}_t$ from which the ego-motion can be then predicted:

$$\mathbf{c}_t = \mathrm{ConvGRU}(\mathcal{E}_\mathrm{m}(\mathbf{h}_t, \mathbf{h}_{t-1}), \mathbf{c}_{t-1}), \qquad \mathrm{C}_{t-1}^{t} = \mathcal{D}_\mathrm{c}(\mathbf{c}_t). \tag{6}$$

Depth Estimation and Ego-Motion Compensation:. Given the past scene state $\mathbf{s}_{t-1}$, the depth decoder $\mathcal{D}_\mathrm{d}$ computes the depth map $\mathbf{d}_{t-1}$. This scene geometry and the estimated ego-motion C_{t-1}^{t} are used as in Equation (1) to project $\mathbf{s}_{t-1}$ to time t. The resulting ego-warped scene state $\mathbf{s}_t^\mathrm{ego}$ encodes scene contents and geometry after compensation for ego-motion.

Residual Flow Estimation and Object Motion Compensation: These modules model dynamic objects in the scene, such as pedestrians or vehicles, and update the scene state to compensate for the motion of such objects. We jointly process the ego-warped scene state $\mathbf{s}_t^\mathrm{ego}$ and the current image features $\mathbf{h}_t$ with the residual flow decoder $\mathcal{R}_\mathrm{f}$ in order to compute the residual flow $\hat{\mathrm{F}}_{t-1}^{t}$, which represents the pixel displacement of moving objects between consecutive time steps. The ego-warped features $\mathbf{s}_t^\mathrm{ego}$ are then projected into the current time-step by applying the displacement encoded in the residual flow map, followed by bilinear interpolation to obtain valid coordinate values. The resulting features $\mathbf{s}_t^\mathrm{full}$ not only incorporate the motion of dynamic objects in the scene, but can also correct alignment errors between $\mathbf{s}_t^\mathrm{ego}$ and $\mathbf{h}_t$ that might occur due to inaccurate depth or ego-motion estimates.

Feature Fusion: While the previous modules propagate scene features over time, the *feature fusion* module allows MCDS-VSS to combine the projected scene features $\mathbf{s}_t^\mathrm{full}$ with the observed encoded features $\mathbf{h}_t$. This fusion operation is performed by an update gate mask $\mathbf{u} \in [0,1]$, which determines in a data-driven manner for each feature map and spatial location whether one can rely on the current features $\mathbf{h}_t$ or on prior knowledge propagated through the filter $\mathbf{s}_t^\mathrm{full}$. The resulting fused scene features $\mathbf{s}_t$ are part of the next state of the MCDS-VSS filter. More formally, this process can be described as:

$$\mathbf{u} = \sigma(\mathrm{Conv}_\mathrm{s}(\mathbf{s}_t^\mathrm{full}) + \mathrm{Conv}_\mathrm{h}(\mathbf{h}_t) + b,) \tag{7}$$

$$\mathbf{s}_t = \mathbf{u} \odot \mathbf{s}_t^\mathrm{full} + (1-\mathbf{u}) \odot \mathbf{h}_t, \tag{8}$$

where Conv_s and Conv_h are convolution blocks, b a learned bias, and σ the sigmoid function.

Table 1: MCDS-VSS training stages and hyper-parameters.

Stage	Training Goal	Loss Function	LR	# Imgs
1	Segmentation & SSL Geometry	$\mathcal{L}_{\text{Segm}} + \lambda_{\text{D}} \cdot \mathcal{L}_{\text{Depth}}$ (2)	$2 \cdot 10^{-4}$	3
2	Distillation of Object Motion	$\mathcal{L}_{\text{Flow}}$ (5)	$1 \cdot 10^{-4}$	2
3	Ego-Motion Filter	$\mathcal{L}_{\text{Ego}}$ (9)	$8 \cdot 10^{-5}$	6
4	Temporal Integration	$\mathcal{L}_{\text{Segm}} + \lambda_{\text{TC}} \cdot \mathcal{L}_{\text{TC}}$ (10)	$8 \cdot 10^{-5}$	6

3.3 Model Training

MCDS-VSS consists of multiple components addressing different subtasks: depth estimation, ego-motion estimation, ego-motion compensation, object motion estimation, object motion compensation, and feature fusion. Naively training such a model in an end-to-end manner with a video segmentation objective can result in bad local optima, where the model does not learn interpretable representations (e.g. depth or object flow).

To ease the training process, we propose a multi-stage training procedure in which we first train the encoder and decoder modules using image pairs or triplets, and then integrate and train the filter modules using sequences of six frames in order to gather scene context information and improve the segmentation performance and temporal consistency while retaining interpretable representations.

MCDS-VSS undergoes a four-stage training process, outlined in Table 1. Initially, as detailed in Section 3.1, MCDS-VSS encoder and decoder modules are jointly trained for self-supervised learning of geometry and ego-motion, as well as for semantic segmentation by minimizing a combination of cross entropy $\mathcal{L}_{\text{Segm}}$ and SSL geometry $\mathcal{L}_{\text{Depth}}$ losses. Following [13], we use image triplets $(\mathbf{x}_{t-\tau}, \mathbf{x}_t, \mathbf{x}_{t+\tau})$, with $\mathbf{x}_t$ being the target image and τ being the temporal distance between source and target frames used during this first training stage. Subsequently, we train the residual flow decoder using image pairs as described in Section 3.1 while keeping the remaining modules frozen. In the third stage, with the goal of improving the temporal continuity of the predicted ego-motion, we train the ego-motion filter and compensation using short video sequences of length T by minimizing the loss function:

$$\mathcal{L}_{\text{Ego}} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_{\text{Photo}}(\hat{\mathbf{x}}_t^{\text{ego}}, \mathbf{x}_t), \tag{9}$$

which enforces the model to compute accurate camera motion estimates in order to align the ego-warped state with the current observations. Finally, in the last training stage we jointly train the feature fusion module and fine-tune the segmentation decoder by minimizing the following loss function:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_{\text{Segm}}(\hat{\mathbf{y}}_t, \mathbf{y}_t) + \lambda_{\text{TC}} \cdot \mathcal{L}_{\text{TC}}(\tilde{\mathbf{y}}_t, \hat{\mathbf{y}}_t), \tag{10}$$

where $\mathcal{L}_{\text{Segm}}$ is the cross entropy loss function and $\mathcal{L}_{\text{TC}}$ is a temporal consistency regularizer that enforces the segmentation $\tilde{\mathbf{y}}_t$ computed by decoding $\mathbf{s}_t^{\text{full}}$ to be close to the actual predicted segmentation maps $\hat{\mathbf{y}}_t$.

Table 2: Comparison of image and video segmentation models on the Cityscapes validation set using small (left) and larger (right) backbones. We evaluate the segmentation accuracy (mIoU) and temporal consistency (TC) of the models. Best two results are highlighted in boldface and underlined, respectively.

Model	Backbone	Cityscapes mIoU↑	Cityscapes TC↑
DeepLabV3+ [6]	ResNet18	75.2	69.8
Accel [20]	ResNet18	72.1	70.3
SKD [26]	ResNet18	74.5	68.2
ETC [27]	PSPNet18	73.1	70.6
ETC [27]	MobileNetV2	73.9	69.9
TCNet [43]	ResNet18	62.2	72.1
TDNet [18]	BiSeNet18	75.0	70.2
TDNet [18]	PSPNet18	<u>76.8</u>	70.4
STT [25]	BiSeNet18	75.8	71.4
STT [25]	ResNet18	**77.3**	<u>73.0</u>
MCDS-VSS (ours)	ResNet18	75.1	**74.5**

Model	Backbone	Cityscapes mIoU↑	Cityscapes TC↑
HRNetV2 [48]	HRNetV2	76.3	70.6
Accel [20]	ResNet50	74.2	-
ETC [27]	HRNetV2	76.4	70.1
ETC [27]	ResNet50	**77.9**	72.3
AuxAdapt [53]	HRNetV2	76.6	**75.3**
PC [54]	HRNetV2	76.4	71.2
TCNet [43]	HRNetV2	72.7	<u>74.7</u>
STT [25]	BiSeNet34	<u>77.3</u>	72.0
MCDS-VSS (ours)	HRNetV2	77.1	**75.3**

4 Experimental Evaluation

4.1 Experiment Setup

Dataset: We evaluate MCDS-VSS on the Cityscapes [8] dataset, which contains 5,000 automotive video sequences recorded in 50 German cities. Each sequence contains 30 images of size 1024×2048, where only the 20th frame is annotated. This dataset is a good benchmark for our model, since it contains real-world dynamic scenes recorded from a moving vehicle. We augment the data using color jittering, mirroring and random cropping.

Evaluation Metrics: We evaluate the segmentation performance and temporal consistency of our model. The performance is evaluated using the mean Intersection-over-Union (mIoU). Following [27], we measure the temporal consistency (TC) of our predicted segmentation maps by computing the mean flow warping error between every two neighboring frames. Our results are computed using single-scale testing on the full image resolution.

Implementation Details: We train two MCDS-VSS variants using distinct image encoder and segmentation decoder architectures. Namely, a small variant based on DeepLabV3+ [6] with ResNet18 [16] backbone, and a larger variant based on HRNetV2 [48]. The depth and pose decoders closely follow [13], which output inverse depth maps and a 6-dimensional vector containing the camera translation and Euler angles, respectively. Finally, our residual flow decoder is a lightweight version of RAFT [42], for which, to integrate into our filter, we replace the context and feature encoders with a single convolutional block. We emphasize that MCDS-VSS is architecture-agnostic and could be implemented with different model designs. Further implementation details are provided in the supplementary material.

4.2 Comparison with Existing Methods

In Table 2, we quantitatively compare MCDS-VSS with several existing image and video segmentation models using small (left) and larger (right) backbones. For both variants, MCDS-VSS achieves the highest temporal consistency among all compared methods, while retaining a competitive segmentation performance. Furthermore, in contrast to other approaches aiming to improve the TC of a segmentation model, e.g. TCNet [43], MCDS-VSS

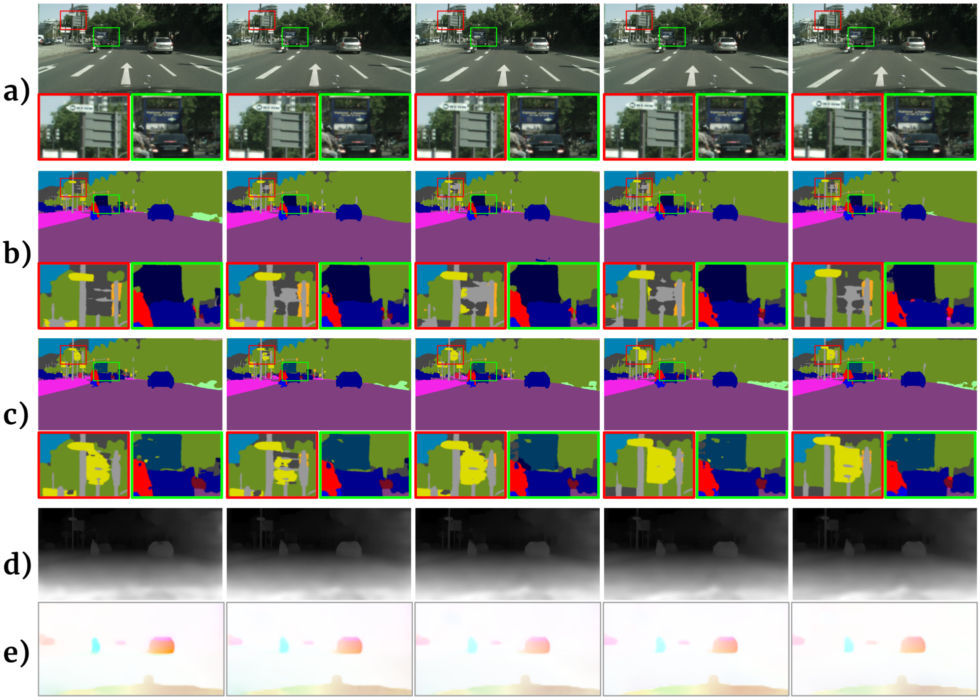


Figure 3: Qualitative evaluation on a validation sequence of five frames. **a)** Input frames, **b)** HRNetV2, **c)** MCDS-VSS (ours), **d)** Estimated scene depth, **e)** Estimated residual flow. We highlight areas of the segmentation masks where MCDS-VSS obtains visibly more accurate and temporally consistent segmentations, such as the traffic signs or the bus, which HRNetV2 mislabels as truck.

does not sacrifice segmentation performance in order to improve the temporal consistency, outperforming multiple VSS models for both backbone variants.

In Figure 3, we show a qualitative result comparing MCDS-VSS with the HRNetV2 baseline on a validation sequence of five frames. Whereas the baseline mislabels the bus as a truck and outputs inconsistent segmentation labels on certain regions such as the traffic signs, our method achieves more accurate and temporally consistent segmentations, predicting more stable semantic labels across video frames. Furthermore, we show MCDS-VSS interpretable intermediate representations, such as the estimated scene depth and residual optical flow, which encodes the movement of the vehicles in the scene, as well as corrections for the hood of the ego-vehicle. Further experiments and visualizations can be found in the supplementary material.

4.3 Ablation Study

To understand the effectiveness of MCDS-VSS, we ablate our filter design and measure the contribution of different steps in our training process.

Filter Design: Given the same DeepLabV3+ model trained with the SSL procedure described in Section 3.1, we compare MCDS-VSS with different filter designs, including unstructured RNNs (ConvGRU) [32, 38], optical-flow based filters [10] (flow-only), and a

Table 3: Comparison of various filter designs. We highlight the diff. to baseline.

Model	Results mIoU↑	TC↑
ResNet18 + SSL	74.76	70.73
+ ConvGRU [38]	73.37 (**-1.39**)	69.64 (**-1.09**)
+ Flow-Only [10]	74.95 (**+0.19**)	73.19 (**+2.46**)
+ Geom-Only [45]	74.90 (**+0.14**)	73.80 (**+3.07**)
+ MCDS-VSS	**75.07** (**+0.31**)	**74.53** (**+3.80**)

Table 4: Effect of SSL geometry & motion and MCDS-VSS. We highlight the diff. to baseline.

Model	Results mIoU↑	TC↑
ResNet18	75.17	69.89
+ SSL Geom. & Motion	74.76 (**-0.44**)	70.73 (**+0.91**)
+ MCDS-VSS	**75.07** (**-0.10**)	**74.53** (**+4.64**)
HRNetV2	76.32	70.58
+ SSL Geom. & Motion	76.45 (**+0.13**)	71.65 (**+1.07**)
+ MCDS-VSS	**77.14** (**+0.82**)	**75.34** (**+4.76**)

MCDS-VSS variant modeling only the static scene features (geom-only) [45]. The results, reported in Table 3, show that filter designs that project scene features using geometry and motion representations outperform the ConvGRU, which learns to model the video dynamics solely from data. Furthermore, MCDS-VSS, which decouples the modeling of static and dynamic scene features, achieves the best segmentation performance and temporal consistency among the compared filter designs.

Model Ablation: In Table 4 we measure the effect that our joint training procedure of semantic segmentation and SSL depth and ego-motion, as well as the MCDS-VSS filter have on the segmentation performance and temporal consistency. For two different segmentation models, i.e. DeepLabV3+ with a ResNet18 backbone and HRNetV2, we compare the results after each training stage with those of the model trained for image segmentation only. First, we note that jointly learning semantic segmentation with SSL depth and ego-motion estimation improves the temporal consistency without significantly compromising the segmentation performance. We argue that the joint training procedure allows the model to encode the input frames into more robust geometry-aware representations. Finally, the MCDS-VSS filter significantly improves the temporal consistency (>4.6% w.r.t base model), while almost matching the segmentation performance of the base DeepLabV3+ model, and even outperforming HRNetV2.

5 Conclusion

We proposed MCDS-VSS, a structured recurrent model for VSS, which learns in a self-supervised manner to estimate scene geometry and camera ego-motion. It also estimates the motion of external objects and leverages these representations to improve the temporal consistency of a semantic segmentation model without sacrificing segmentation performance. MCDS-VSS follows a prediction-fusion approach in which scene geometry and camera motion are first used to compensate for ego-motion, then residual flow is used to compensate the motion of dynamic objects, and finally the projected features are fused with the current observations in order to obtain a temporally consistent representation of the scene. In our experiments, we showed that MCDS-VSS outperforms multiple VSS baselines on Cityscapes—achieving superior segmentation temporal consistency and parsing the scene into human-interpretable representations, such as depth, ego-motion and object flow.

Acknowledgment

This work was funded by grant BE 2556/16-2 (Research Unit FOR 2535 Anticipating Human Behavior) of the German Research Foundation (DFG).

References

[1] Ali Athar, Enxu Li, Sergio Casas, and Raquel Urtasun. 4D-Former: Multimodal 4D panoptic segmentation. In *Conference on Robot Learning (CoRL)*, pages 2151–2164, 2023.

[2] Razieh Kaviani Baghbaderani, Yuanxin Li, Shuangquan Wang, and Hairong Qi. Temporally-consistent video semantic segmentation with bidirectional occlusion-guided feature propagation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 685–695, 2024.

[3] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *International Conference on Learning Representations (ICLR)*, 2016.

[4] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision (IJCV)*, 129(9):2548–2564, 2021.

[5] Helin Cao and Sven Behnke. SLCF-Net: Sequential LiDAR-Camera Fusion for Semantic Scene Completion using a 3D Recurrent U-Net. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[7] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. Spatial information guided convolution for real-time RGBD semantic segmentation. *IEEE Transactions on Image Processing*, 30:2313–2324, 2021.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[9] Mingyu Ding, Zhe Wang, Bolei Zhou, Jianping Shi, Zhiwu Lu, and Ping Luo. Every frame counts: Joint learning of video segmentation and optical flow. In *Conference on Artificial Intelligence (AAAI)*, volume 34, pages 10713–10720, 2020.

[10] Raghudeep Gadde, Varun Jampani, and Peter V Gehler. Semantic video CNNs through representation warping. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[11] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision (ECCV)*, pages 740–756, 2016.

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 270–279, 2017.

[13] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3828–3838, 2019.

[14] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8977–8986, 2019.

[15] Vitor Guizilini, Jie Li, Rares Ambrus, and Adrien Gaidon. Geometric unsupervised domain adaptation for semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8537–8547, 2021.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Koring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11130–11140, 2021.

[18] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8818–8827, 2020.

[19] Junhwa Hur and Stefan Roth. Joint optical flow and temporally consistent semantic segmentation. In *European Conference on Computer Vision Workshops (ECCVw)*, pages 163–177. Springer, 2016.

[20] Samvit Jain, Xin Wang, and Joseph E Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8866–8875, 2019.

[21] Abhijit Kundu, Xiaoqi Yin, Alireza Fathi, David Ross, Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru. Virtual multi-view fusion for 3D semantic segmentation. In *European Conference on Computer Vision (ECCV)*, pages 518–535, 2020.

[22] Yevhen Kuznietsov, Marc Proesmans, and Luc Van Gool. Towards unsupervised online domain adaptation for semantic segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 261–271, 2022.

[23] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14763–14772, 2023.

[24] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon. Learning monocular depth in dynamic scenes via instance-aware projection consistency. In *Conference on Artificial Intelligence (AAAI)*, pages 1863–1872, 2021.

[25] Jiangtong Li, Wentao Wang, Junjie Chen, Li Niu, Jianlou Si, Chen Qian, and Liqing Zhang. Video semantic segmentation via sparse temporal transformer. In *29th ACM International Conference on Multimedia (MM)*, pages 59–68, 2021.

[26] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2604–2613, 2019.

[27] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *European Conference on Computer Vision (ECCV)*, pages 352–368, 2020.

[28] Lingni Ma, Jörg Stückler, Christian Kerl, and Daniel Cremers. Multi-view deep learning for consistent semantic mapping with RGB-D cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 598–605, 2017.

[29] Jelena Novosel, Prashanth Viswanath, and Bruno Arsenali. Boosting semantic segmentation with multi-task self-supervised learning for autonomous driving applications. In *NeurIPS-Workshops*, volume 3, 2019.

[30] Andra Petrovai and Sergiu Nedevschi. Monodvps: A self-supervised monocular depth estimation approach to depth-aware video panoptic segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3077–3086, 2023.

[31] Andreas Pfeuffer and Klaus Dietmayer. Separable convolutional lstms for faster video segmentation. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 1072–1078, 2019.

[32] Andreas Pfeuffer, Karina Schulz, and Klaus Dietmayer. Semantic segmentation of video sequences with convolutional lstms. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1441–1447, 2019.

[33] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. ViP-DeepLab: Learning visual perception with depth-aware video panoptic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3997–4008, 2021.

[34] Radu Alexandru Rosu, Jan Quenzel, and Sven Behnke. Semi-supervised semantic mapping through label propagation with semantic texture meshes. *International Journal of Computer Vision (IJCV)*, 128(5):1220–1238, 2020.

[35] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient RGB-D semantic segmentation for indoor scene analysis. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531, 2021.

[36] Evan Shelhamer, Kate Rakelly, Judy Hoffman, and Trevor Darrell. Clockwork convnets for video semantic segmentation. In *European Conference on Computer Vision Workshops (ECCVw)*, pages 852–868, 2016.

[37] Hanyu Shi, Guosheng Lin, Hao Wang, Tzu-Yi Hung, and Zhenhua Wang. SpSequenceNet: Semantic segmentation network on 4D point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4574–4583, 2020.

[38] Mennatullah Siam, Sepehr Valipour, Martin Jagersand, and Nilanjan Ray. Convolutional gated recurrent networks for video segmentation. In *IEEE International Conference on Image Processing (ICIP)*, pages 3090–3094, 2017.

[39] Mennatullah Siam, Mostafa Gamal, Moemen Abdel-Razek, Senthil Yogamani, Martin Jagersand, and Hong Zhang. A comparative study of real-time semantic segmentation for autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRw)*, pages 587–597, 2018.

[40] Guolei Sun, Yun Liu, Hao Tang, Ajad Chhatkuli, Le Zhang, and Luc Van Gool. Mining relations among cross-frame affinities for video semantic segmentation. In *European Conference on Computer Vision (ECCV)*, pages 522–539, 2022.

[41] 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 (TPAMI)*, 2024.

[42] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)*, pages 402–419, 2020.

[43] Serin Varghese, Sharat Gujamagadi, Marvin Klingner, Nikhil Kapoor, Andreas Bar, Jan David Schneider, Kira Maag, Peter Schlicht, Fabian Huger, and Tim Fingscheidt. An unsupervised temporal consistency (TC) loss to improve the performance of semantic segmentation networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRw)*, pages 12–20, 2021.

[44] Angel Villar-Corrales, Ani Karapetyan, Andreas Boltres, and Sven Behnke. MSPred: Video prediction at multiple spatio-temporal scales with hierarchical recurrent networks. In *British Machine Vision Conference (BMVC)*, 2022.

[45] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Functionally modular and interpretable temporal filtering for robust segmentation. In *British Machine Vision Conference (BMVC)*, page 282, 2018.

[46] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2022–2030, 2018.

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

[48] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[49] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 8515–8525, 2021.

[50] Yu-Syuan Xu, Tsu-Jui Fu, Hsuan-Kung Yang, and Chun-Yi Lee. Dynamic video segmentation network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6556–6565, 2018.

[51] Haobo Yuan, Xiangtai Li, Yibo Yang, Guangliang Cheng, Jing Zhang, Yunhai Tong, Lefei Zhang, and Dacheng Tao. Polyphonicformer: Unified query learning for depth-aware video panoptic segmentation. In *European Conference on Computer Vision (ECCV)*, pages 582–599. Springer, 2022.

[52] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. In *14th Asian Conference on Computer Vision (ACCV)*, pages 298–313. Springer, 2019.

[53] Yizhe Zhang, Shubhankar Borse, Hong Cai, and Fatih Porikli. AuxAdapt: Stable and efficient test-time adaptation for temporally consistent video semantic segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2339–2348, 2022.

[54] Yizhe Zhang, Shubhankar Borse, Hong Cai, Ying Wang, Ning Bi, Xiaoyun Jiang, and Fatih Porikli. Perceptual consistency in video segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2564–2573, 2022.

[55] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[56] Tianfei Zhou, Fatih Porikli, David J Crandall, Luc Van Gool, and Wenguan Wang. A survey on deep learning technique for video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(6):7099–7122, 2022.

[57] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2017.

[58] Minghan Zhu, Shizhong Han, Hong Cai, Shubhankar Borse, Maani Ghaffari, and Fatih Porikli. 4D panoptic segmentation as invariant and equivariant field prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22488–22498, 2023.

[59] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2349–2358, 2017.