Unsupervised Temporal Consistency Metric for Video Segmentation in Highly-Automated Driving

Serin Varghese[1,2*] Yasin Bayzidi[1*] Andreas Bär[2] Nikhil Kapoor[1] Sounak Lahiri[1]
Jan David Schneider[1] Nico Schmidt[1] Peter Schlicht[1] Fabian Hüger[1] Tim Fingscheidt[2]

{john.serin.varghese, yasin.bayzidi, nikhil.kapoor, sounak.lahiri, jan.david.schneider, nico.maurice.schmidt, peter.schlicht, fabian.hueger}@volkswagen.de

{s.varghese, andreas.baer, t.fingscheidt}@tu-bs.de

[1]Volkswagen AG [2]Technische Universität Braunschweig

Abstract

Commonly used metrics to evaluate semantic segmentation such as mean intersection over union (mIoU) do not incorporate temporal consistency. A straightforward extension of existing metrics towards evaluating the consistency of segmentation of video sequences does not exist, since labelled videos are rare and very expensive to obtain. For safety-critical applications such as highly automated driving, there is, however, a need for a metric that measures such temporal consistency of video segmentation networks to possibly support safety requirements. In this paper, (a) we introduce a metric which does not require segmentation labels for measuring the stability of the predictions of segmentation networks over a series of images; (b) we perform an in-depth analysis of the proposed metric and observe strong correlations to the supervised mIoU metric; (c) we perform an evaluation of five state-of-the-art networks for semantic segmentation of varying complexities and architectures evaluated on two public datasets, namely, Cityscapes and CamVid. Finally, we perform timing evaluations and propose the use of the metric as either an online observer for identification of possibly unstable segmentation predictions, or as an offline method to evaluate or to improve semantic segmentation networks , e.g., by selecting additional training data with critical temporal consistency.

1. Introduction

The success of deep neural networks (DNNs) in computer vision and pattern recognition tasks makes them a promising technology for a wide variety of applications, including highly-automated driving perception systems.

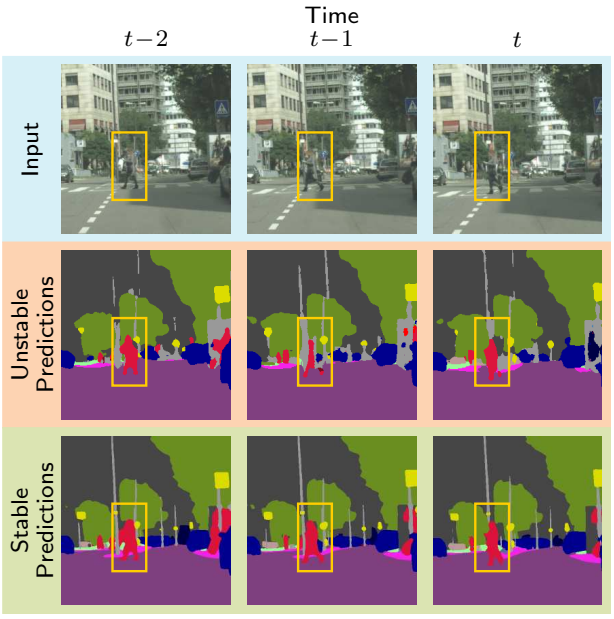

Figure 1: **Examples of stable and unstable predictions** of semantic segmentation networks on videos. The yellow boxes highlight the segmentation area of interest in the frames. *Top*: Left to right are the frames at discrete time instances $t-2$, $t-1$, and t of a video sequence that are fed into a semantic segmentation network. *Middle*: The predictions of the pedestrian are unstable, i.e., they are not consistent across time (both $t-2 \rightarrow t-1$, and $t-1 \rightarrow t$). *Bottom*: The predictions are more stable over time. Our new metric helps in evaluating these instantaneous inconsistencies in the predictions.

*These authors contributed equally.

There have been a lot of improvements in vision-related tasks such as image classification [18, 20, 25, 43, 44, 45], object detection [16, 17, 38, 39], and semantic segmentation [2, 3, 5, 28, 33, 41]. We focus on semantic segmentation networks that perform dense pixel-wise classification of input images into a set of predefined semantic classes. For both training and evaluation, ground truth labels are necessary, which are usually considered expensive to obtain. In order to assess the quality of these semantic segmentation models, metrics such as mean intersection over union (mIoU) are commonly used. This gives an estimate of how accurately the output pixels are assigned to their respective semantic classes on a pre-defined test set of images. The test set is designed such that it consists of individual frames which are considered representative of a diverse set of scenes and objects seen in the real world [15]. However, in the real world, cameras usually capture *video* sequences and not individual frames over different time intervals. Since existing quality metrics such as mIoU operate on individual frames, they do not capture any temporal notions and hence offer only limited insights. We argue that in order to better assess models working on videos, we need metrics that capture additional temporal characteristics in addition to simple pixel-level accuracy measures. One such example of a temporal characteristic would be continuity of objects in consecutive frames that we also demonstrate in Fig. 1. Assuming a video sequence of sufficiently high frame rate, this means that objects usually do not appear or disappear suddenly in consecutive frames, except from behind occlusions. To the best of our knowledge, there are no existing metrics that capture such notions.

At present, evaluating video semantic segmentation is considered challenging due to two reasons. First, there is a severe lack of publicly available *video* semantic segmentation datasets offering a large number of *consecutive* high-quality annotated images. Labelling long video sequences is expensive and does not produce major improvements for the semantic segmentation networks as the diversity of the dataset is not large enough for the network to generalise from during training. Second, since existing quality metrics do not capture any temporal characteristics, they offer only limited insights. Additionally, certain methods [24, 42] exploit the temporal consistency within videos to improve the quality of semantic segmentation networks. However, these approaches only perform evaluations on the improvement of accuracy and do not measure the effects of the methods on temporal consistency of their predictions.

In this paper, we address the abovementioned challenges by proposing a new evaluation metric that is suitable for assessing models on video sequences. This temporal consistency metric captures the notion of *temporal consistency*, which we define as the measurement of the sudden appearance and disappearance of objects in consecutive frames. In addition, this proposed metric is fully *unsupervised* by nature, which means that it does not require any expensive labelling procedures.

Our new metric could be used as an *observer* in the automated vehicle, a system or evaluation that runs in parallel to primary perception modules in the vehicle to identify instances, where there might be a possibility of failure [4, 6, 46]. In the case of detection of sudden instability, this observer could possibly give an additional input which could be combined with rule-based systems to avoid a mishap.

As one of the major challenges of using DNNs for highly automated driving is to ensure safety requirements of neural networks, defining methods and metrics for measuring their robustness-oriented traits has become an active and important research field. The authors propose that *stable detection of objects over time* could be one of the safety criteria for automated driving, and a metric to evaluate this might be helpful.

In this paper, our major contributions are as follows: We introduce a novel metric to measure the consistency of the predictions of semantic segmentation models, we show that our unsupervised metric has a strong correlation with the supervised intersection over union metric, and finally we perform inference time evaluations and show that our metric could be used as an observer in the vehicle. To the best of our knowledge, this is the first time detailed quantitative evaluation of the temporal consistency of semantic segmentation has been made possible.

This paper is structured as follows: Section 2 reviews the related work. In Section 3, we describe our intuition and explain our metric to measure the temporal consistency of semantic segmentation predictions. In Section 4, we present our results and observations. Finally, we conclude in Section 5.

2. Related Work

In this section, we discuss the related work in the field of semantic segmentation evaluation including temporal consistency measurement approaches and optical flow methods.

Semantic segmentation evaluation methods can be broadly classified in two ways, namely *supervised* and *unsupervised* evaluation [21]. Supervised methods depend on the existence of labelled ground truth, while unsupervised methods do not need this. Everingham *et al*. [12] introduced intersection over union (IoU) between the predicted segmentation mask and the labelled ground truth. Martin *et al.* [30] defined boundary precision recall, which evaluates the semantic segmentation based on the detection of the boundaries. For evaluating video

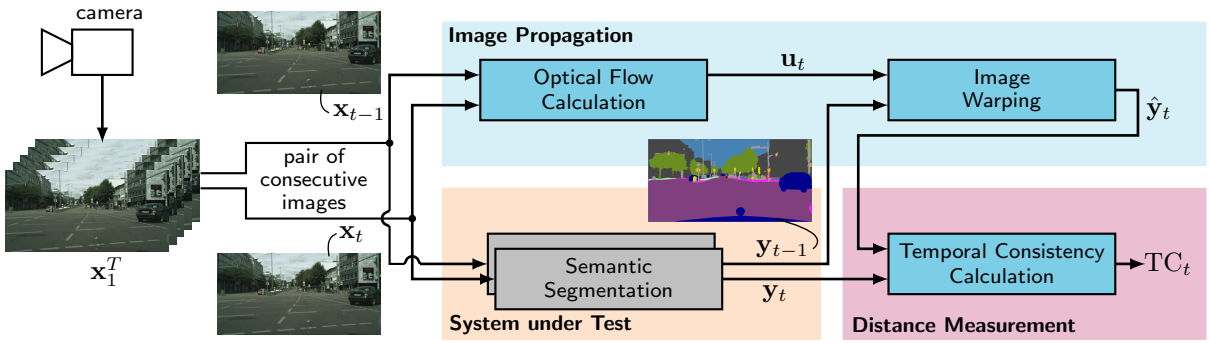

Figure 2: **Novel temporal consistency metric** TC_t (4) of the dense predictions from a semantic segmentation model.

sequences, Galasso *et al*. [13] introduce volume precision recall (VPR), which creates spatio-temporal volumes of the predictions of semantic segmentation. Unsupervised methods do not depend on labelled ground truth for evaluation. Maag *et al*. [29] propose a method to estimate the reliability of semantic segmentation based on temporal uncertainty measurements. To this end, the meta-classifier is trained on segment-wise, or connected component-wise, metrics which are tracked over time.

Perazzi *et al*. [35] introduce the DAVIS dataset for video segmentation, define temporal stability, and measure the error by defining *acceptable* and *undesired* changes in two consecutive frames. To do so, polygons of shape contours are extracted from the predicted segmentation and fed into a shape context descriptor (SCD). Afterwards, the transformation of masks from one frame to the other are calculated using SCD distances, which are minimized using dynamic time warping correlation calculation. Finally, if a smooth transformation is observed in more than two consecutive frames, the predictions are considered as *stable*. However, this metric is discarded by the same authors in [7, 8, 37], as it is found sensitive to occlusions.

Kundu *et al*. [26] and Nilson *et al*. [32] use a consistency metric to capture the concept of temporal consistency. For this purpose, they use the large displacement optical flow (LDOF) defined by Sundaram *et al*. [31]. With LDOF, long-term tracks are computed in the video which are used for the definition of consistency. A track is termed as consistent if all the pixels are assigned the same label by the semantic segmentation network. This consistency measure can only be calculated over an entire video sequence and does not provide instantaneous consistency of video segmentation models at any time instant t, which, however, is required for its use in observer applications. The consistency method [26] will serve as baseline for us.

Optical flow algorithms estimate the displacement of pixels from one video frame to another. Gunnar *et al*. [14] proposed a multi-scale optical flow estimation, where the two images are downsampled multiple times, shaping a pyramid, while the displacement of the pixels is calculated for each level of this pyramid. This way, small displacements are detected in lower levels of the pyramid, and the large displacements are estimated in the higher levels, where the images are much smaller. Sundaram *et al*. [31] propose a forward-backward method, where each displacement vector is calculated twice; once from time $t-1$ to t (forward) and once backward from t to $t-1$. This way, if a pixel gets a similar displacement vector in both directions, it is considered to be a correct displacement vector, otherwise the pixel might be occluded in the later frame and can be eliminated. Ilg *et al*. [22] introduce FlowNet2 as a newer version of FlowNet [11], which is a neural network trained to estimate the optical flow. It has been shown that this approach can generate comparative and sometimes even better results than the conventional optical flow methods. TVL1 [36] is an extension of the well-known Horn-Schunck method [19], where the linear flow estimation is replaced by a non-linear term, which allows to account for discontinuities. It has been shown that the discontinuities of flows are covered by this method and it is more robust to noise than the original approach.

For estimating flow vectors for our metric, we use the optical flow approach from Gunnar *et al*. [14], which has a good trade-off between accuracy and run-time. We also use the state-of-the-art approaches mentioned above to perform an ablation study to study the effect of various optical flow algorithms.

3. Method

In this section, we introduce our metric for semantic segmentation models. We term this as temporal consistency TC, which measures the stability of semantic segmentation predictions on video sequences. Driven by the motivation given above, we introduce the intuition behind our metric and explain in detail the steps to calculate TC.

Intuition: Before introducing our metric, we first briefly

introduce the idea of temporal consistency. A well-working semantic segmentation network should produce *similar* predictions between two frames in a video stream. *Similar*, because in a video sequence there is little variation between consecutive frames, depending on the frame rate of the camera. The variation depends on two aspects: the movement of the objects in the video, and the translational and rotational movement of the camera. We expect the predictions of video segmentations to be stable if we can compensate for these variations. Taking a simple difference between two predictions of consecutive frames does not take into consideration these variations. However, optical flow approaches can be used to model the *movement* of pixels between two images. This maps both, the movement of the objects in the video and also the movement of the camera. By such accurate modelling using flow functions, we can warp the frame at time $t-1$ to the frame at time t, which then helps in calculating the instantaneous stability of the network's predictions.

In order to introduce the temporal consistency for semantic segmentation predictions, we follow Fig. 2, which displays all the steps involved. In our more detailed explanation on how the temporal consistency TC is calculated, we refer to the blocks in Fig. 2. The procedure of calculating the temporal consistency of a semantic segmentation is as follows:

Semantic segmentation: The system under test is a semantic segmentation model $\mathfrak{F}$, whose temporal consistency has to be measured. We first compute the semantic segmentation predictions of two consecutive sequential images. We define $\mathbf{x} \in \mathbb{G}^{H\times W\times C}$ to be an image of height H, width W, number of color channels $C = 3$, and pixel intensities $\mathbb{G} = \{0, 1, 2, \ldots, 255\}$. Let $\mathbf{x}_1^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ be the unlabelled video sequence of consecutive images of length T, and $\mathbf{x}_t$ be an image of this sequence at discrete time instant $t \in \mathcal{T} = \{1, 2, \ldots, T\}$. With $\mathbf{y}_t = \mathfrak{F}(\mathbf{x}_t)$ we denote the prediction of the semantic segmentation network, where $\mathbf{y}_t \in \mathcal{S}^{H\times W}$ and $\mathcal{S} = \{1, 2, \ldots, S\}$ where S is the number of classes in the dataset. Thus, $\mathbf{y}_{t-1}$ and $\mathbf{y}_t$ are the predictions of consecutive images $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$, respectively.

Optical flow calculation: We use optical flow functions to capture the apparent motion within the video sequence. Optical flow estimates the displacement of each pixel between the consecutive frames $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$. The computed optical flow between $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$ is defined as a tensor $\mathbf{u}_t \in \mathcal{U}^{H\times W}$ following [31], where $\mathcal{U}$ is the set of two-dimensional pixel-wise displacements (u, v), representing the coordinate-wise shift of each pixel from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$. We use these pixel-wise displacements to apply the same shift to the pixel coordinates of the segmentation output $\mathbf{y}_{t-1}$ as in [23]. This way, we generate an expected segmentation output $\hat{\mathbf{y}}_t$ based on the calculated flow, representing pixel-wise shifts from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$.

Image warping: Using the derived optical flow tensor $\mathbf{u}_t$ calculated between $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$, the prediction of the semantic segmentation network $\mathbf{y}_{t-1}$ is warped from time t–1 to time t. To do this, we first define pixel coordinates for an image as tensor $\mathbf{p} \in \mathcal{P}^{H\times W}$, where $\mathcal{P}$ is the set of pixel positions (or index pairs) (i, j) with $i \in \{1, \ldots, H\}$ and $j \in \{1, \ldots, W\}$. Tensor $\mathbf{p}$ thus only contains the pixel-wise coordinates of a pixel in an image and does not carry any information about pixel intensity values. Now we can add the pixel-wise displacement vectors $\mathbf{u}_t$ to the original pixel positions $\mathbf{p}$ to receive a tensor

$$\mathbf{p}_{t-1\to t} = \mathbf{p} + \mathbf{u}_t, \tag{1}$$

which provides the new pixel coordinates. Subsequently, we apply $\mathbf{p}_{t-1\to t}$ to the segmentation output $\mathbf{y}_{t-1}$. As the pixel coordinates $\mathbf{p}_{t-1\to t}$ are non-integer numbers, we use nearest neighbour sampling $\mathrm{nearest}()$ as described by [23] to obtain valid integer coordinates for mapping of $\mathbf{y}_{t-1}$ to the flow-based estimate $\hat{\mathbf{y}}_t$ using $\mathbf{p}_{t-1\to t}$. That is we warp $\mathbf{y}_{t-1}$ to $\hat{\mathbf{y}}_t$ by

$$\hat{\mathbf{y}}_t = \mathrm{nearest}(\mathbf{y}_{t-1}, \mathbf{p}_{t-1\to t}). \tag{2}$$

Accordingly, $\hat{\mathbf{y}}_t$ is the *expected* prediction at time t based on the optical flow and conditioned on the change in the pair of inputs $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$, which compensates for the movement of the camera and the objects in the consecutive frames.

Temporal consistency calculation: Ideally, for a good semantic segmentation model, the distance between the network output $\mathbf{y}_t$ and the prediction based on the optical flow $\hat{\mathbf{y}}_t$ should be small. We now define the instantaneous temporal consistency of semantic segmentation predictions as

$$\mathrm{TC}_t = \mathrm{mIoU}(\mathbf{y}_t, \hat{\mathbf{y}}_t), \tag{3}$$

calculating the mean intersection over union [12] between $\mathbf{y}_t$ and $\hat{\mathbf{y}}_t$, where $\mathrm{mIoU} = 1$ indicates that both completely overlap. This calculation can be done on all pairs of consecutive images in a video sequence. Subsequently, if a single metric value is desired, the mean temporal consistency is obtained by

$$\mathrm{mTC} = \frac{1}{T'-1}\sum_{t=2}^{T'} \mathrm{TC}_t, \tag{4}$$

where T' is the number of frames in the test set. The mean temporal consistency metric mTC therefore indicates the stability of the predictions of a semantic segmentation model by motion flow calculations, given a test video sequence, without requiring labels.

Networks	Backbone Architecture	FLOPs (billion)	Model Size [MB]	mIoU [%]
ICNet[1] [48]	PSPNet	58.27	26.8	70.6
ERFNet[2] [40]	own	214.35	8.2	71.7
DeepLab v3+[3] [9]	Xception65	2778.24	165.7	77.8
ResNet38-ASPP[4] [5]	WideResNet	11056.13	543.3	77.8
HRNetv2[5] [47]	ResNet50	748.78	263.1	81.1

Table 1: **Overview of semantic segmentation models** used for evaluation. The networks are taken from their respective github repositories, if available, where we use the models pretrained on Cityscapes. Numbers are provided in billion FLOPs computed for 1024×2048 input images, model size [MByte], and mIoU [%] on the validation set. Note that all the model parameters are saved as 32 bit float numbers.

Dataset	Seq ID	Model	C26 [26] [%]	mTC (4) [%]	mIoU (6) [%]	r (7)
Cityscapes	00	ICNet	87.26	72.72	52.59*	0.71
		ERFNet	85.90	67.52	55.09*	0.86
		DeepLabv3+	86.82	72.67	64.50*	0.95
		ResNet38A	**89.58**	75.95	69.40*	0.77
		HRNetv2	89.56	**79.15**	100	-
	01	ICNet	89.46	72.07	55.13*	0.81
		ERFNet	87.31	66.17	56.35*	0.82
		DeepLabv3+	88.47	72.54	67.87*	0.92
		ResNet38A	90.71	75.79	70.80*	0.85
		HRNetv2	**90.81**	**77.28**	100	-
	02	ICNet	86.12	70.53	50.74*	0.54
		ERFNet	85.73	67.25	56.20*	0.73
		DeepLabv3+	86.51	73.09	63.18*	0.72
		ResNet38A	88.36	75.14	64.43*	0.63
		HRNetv2	**88.42**	**78.33**	100	-
CamVid	16E5	ICNet	85.81	70.53	57.13	0.96
		ERFNet	69.96	87.96	80.06	0.74
		DeepLabv3+	82.19	80.56	56.98	0.97
		ResNet38A	90.24	**88.09**	77.60	0.82
		HRNetv2	**90.85**	75.67	65.81	0.72

Table 2: **Temporal coherence results (C26 [26], novel mTC (4) with optical flow method adopted from [31]), mIoU (6) and Pearson correlation r (7) between TC_t (3) and $mIoU_t$ (5).** The numbers are reported for the semantic segmentation models in Tab. 1 evaluated on the sequences of the Cityscapes dataset and CamVid dataset (see Section 4.2). Stars (*) indicate that the mIoU value is calculated using pseudo-ground truth from HRNetv2. Best temporal consistencies are shown in **bold**.

4. Experiments and Results

In this section, we describe the datasets and the semantic segmentation networks we use for evaluation. We also present the evaluation of the mean temporal consistency metric mTC in comparison to existing works. Finally, we study the correlation of the unsupervised consistency metric with the intersection over union metric, which is particularly relevant for the observer use case in the vehicle.

4.1. Semantic Segmentation Networks

Table 1 provides an overview of the networks that we have used for evaluating their temporal consistency on video segmentation. The semantic segmentation networks are chosen such that they have varying backbones, architectures, and model sizes. Efficient semantic segmentation networks such as ICNet [48][1] and ERFNet [40][2] are used. We also use bigger models such as DeepLab v3+ [9][3], ResNet38-ASPP [27][4], and HRNetv2 [47][5]. ICNet is a fast real-time network which is ideal for edge applications, whereas HRNetv2 is more accurate but lacks real-time capability. For ERFNet, "own" indicates that the backbone architecture is a novel architecture that was introduced in [40]. The values of FLOPs are calculated for the Cityscapes images with image dimensions of 1024×2048. This selection of networks helps also in studying the inter-dependence of temporal consistency with inference time and segmentation accuracy.

[1] https://github.com/hellochick/ICNet-tensorflow
[2] https://github.com/Eromera/erfnet
[3] https://github.com/tensorflow/models/tree/master/research/deeplab
[4] We follow the training procedure as given in [27].
[5] https://github.com/HRNet/HRNet-Semantic-Segmentation

4.2. Datasets

In this section, we describe the semantic segmentation datasets we have used to evaluate our metric.

Cityscapes dataset: The Cityscapes dataset contains $5,000$ images from different cities, of which $2,975$ images are used for training, 500 images for validation, and $1,525$ for testing. We utilize the sequential unlabelled demo videos provided within the dataset to evaluate our metric. Similar to Cordts *et al.* [10], we also reduce the 33 classes to 19 relevant classes by excluding classes that are too rare. We evaluate the mean temporal consistency (mTC) on the three unlabelled sequences available. These three sequences, `stuttgart_00`, `stuttgart_01` and `stuttgart_02` are described as sequences 00, 01 and 02 in Table 2. We choose the Cityscapes dataset because of its diversity of highly dynamic objects present in road scenes. This dataset is also a widely used and accepted benchmark for semantic segmentation in general.

CamVid dataset: The CamVid dataset contains 701 images from which we use 367 training images, 100

validation images, and 233 test images. This split is similar to Sturgess *et al*. [34], which will ease comparison with previous works on this dataset. A number of 11 semantic classes are used for evaluation. For the temporal consistency measurement we use the sequence `Seq16E5`, consisting of 101 frames captured at a frequency of 15Hz (16E5 in Table 2). We use this dataset, as this is the widely used road scenes dataset that provides video sequences for evaluation.

4.3. Experiments

In this section, we compare our metric mTC (4) with consistency C26 as defined in [26, 32], where long-term tracks are calculated over the entire sequence, based on the approach from Sundaram *et al*. [31]. For C26, a consistent track is defined as a track, where all the pixels along the track have the same label assigned by the predictions of the semantic segmentation network. Consistency C26 in the end is the percentage of consistent tracks across the video sequence. The comparative results are shown in Table 2. For the Cityscapes dataset, we observe that mTC is highest across all sequences for the powerful HRNetv2, whereas for the CamVid dataset ResNet38-ASPP shows the best mTC. Looking at the C26 metric [26], we observe that it is more or less constant over all the networks and all sequences. However, this cannot be confirmed from the temporal instabilities that are seen in the video segmentations. Fig. 4 shows example predictions of the semantic segmentation networks, where it can be observed that smaller networks tend to have a higher instability than the bigger ones. Our metric mTC captures this instability better as it is calculated for consecutive images (3) and then averaged over the video sequence (4). With respect to computation time, *we observed our metric mTC to be faster than C26 by a factor of about 10*. This, along with its instantaneous option TC_t (3) enables the use of TC_t in real-time applications very much unlike [26].

4.4. Correlation With the mIoU Metric

In this section, we study the correlation of the unsupervised TC_t (3) with the supervised intersection over union metric. The mean intersection over union mIoU_t at time t is defined as

$$\mathrm{mIoU}_t = \frac{1}{S}\sum_{s\in\mathcal{S}} \frac{\mathrm{TP}_t(\mathrm{s})}{\mathrm{TP}_t(\mathrm{s}) + \mathrm{FP}_t(\mathrm{s}) + \mathrm{FN}_t(\mathrm{s})} = \mathrm{mIoU}(\mathbf{y}_t, \bar{\mathbf{y}}_t), \tag{5}$$

where $\mathrm{TP}_t(\mathrm{s})$, $\mathrm{FP}_t(\mathrm{s})$ and $\mathrm{FN}_t(\mathrm{s})$ are the class-specific true positives, false positives, and false negatives, respectively, computed only for segmentation output $\mathbf{y}_t$. The mean intersection over union (mIoU) is defined as

$$\mathrm{mIoU} = \frac{1}{T'}\sum_{t=1}^{T'} \mathrm{mIoU}_t. \tag{6}$$

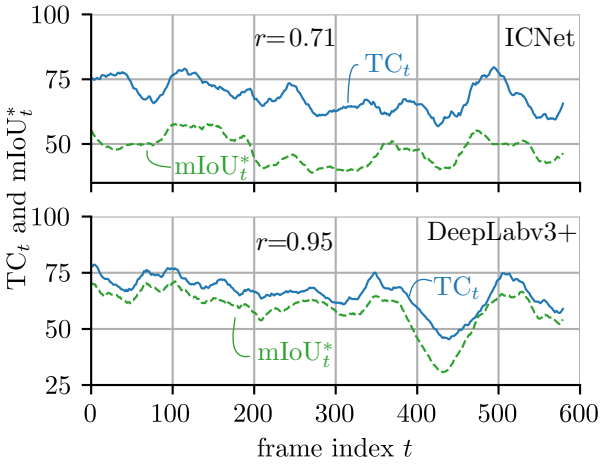

Figure 3: mIoU_t* and TC_t for the ICNet (*top*) and the DeepLabv3+ (*bottom*) model trained on Cityscapes. The evaluation is performed on the `stuttgart_00` sequence. This plot highlights the strong correlation between our proposed temporal consistency TC_t metric and the mean intersection over union metric with pseudo-ground truth mIoU_t*.

As the ground truth labels are not available for the Cityscapes videos, we use the powerful HRNetv2 [47] to generate pseudo-ground truth for these sequences. The calculated mean intersection-over-union is termed mIoU_t*, and its mean over the sequence mIoU* is shown in Table 2. Based on these pseudo-ground truth labels and the predictions of the semantic segmentation networks, the Pearson cross-correlation coefficient [1] is measured between $a_t = \mathrm{mIoU}_t$* and $b_t = \mathrm{TC}_t$, $t \in \mathcal{T}$. Pearson cross-correlation is defined as

$$r_{a,b} = \frac{\sum_{t\in\mathcal{T}}(a_t - \mu_a)(b_t - \mu_b)}{\sqrt{\sum_{t\in\mathcal{T}}(a_t - \mu_a)^2}\sqrt{\sum_{t\in\mathcal{T}}(b_t - \mu_b)^2}} \tag{7}$$

where μ_a and μ_b are the respective means. Here, $r = -1$ indicates perfect anti-correlation, $r = 0$ indicates no (linear) correlation, and $r = 1$ indicates perfect positive correlation between the time series of a and b. Similarly for the CamVid dataset, we use the ground truth to calculate the mean intersection over union at time t and the Pearson cross-correlation is calculated between mIoU_t (5) and TC_t (3). From Table 2 we observe *strong positive correlations for all the networks over all the sequences and datasets*. In Fig. 3, we visualise this correlation between mIoU_t and TC_t for two models on one unlabelled Cityscapes sequence. As the metric by [26] is calculated as the percentage of long-term tracks having the same labels assigned by the network, they provide only one value per sequence. Performing a similar correlation analysis on their metric is therefore not

Method	ICNet [48]		ERFNet [40]		Deeplab v3+ [9]		ResNet38A [5]		HRNetv2 [47]		MSE		Time
	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV	[s]
mTC + [14]	0.72	0.62	0.71	0.72	0.76	0.66	0.79	0.74	0.81	0.64	0.0037	0.0083	1.08
mTC + [36]	0.73	0.63	0.72	0.74	0.77	0.71	0.83	0.75	0.79	0.65	0.0050	0.0103	12.91
mTC + [31]	0.72	0.71	0.67	0.88	0.73	0.81	0.76	0.88	0.78	0.76	**0.0016**	**0.0031**	3.70
mTC + [22]	0.74	0.70	0.71	0.86	0.77	0.79	0.82	0.87	0.83	0.75	0.0031	0.0062	**0.36**

Table 3: Ablation study on the **effect of different optical flow estimation methods on the mIoU temporal consistency metric mTC**. Evaluations are performed on sequences of the Cityscapes (CS) dataset and the CamVid (CV) dataset (see Section. 4.2). The mean squared error (MSE) is calculated pixel-wise between $\mathbf{x}_t$ and the warped image $\hat{\mathbf{x}}_t$ based on the optical flow. A lower MSE value indicates a higher accuracy of the optical flow. Additionally, the average time required to execute an optical flow calculation between two images within the sequence is reported. The best numbers are printed in **bold**.

possible, without significantly modifying their approach to computing stability of predictions.

Due to these strong correlations, our fully unsupervised mTC metric can be used to identify and select additional training data, where we observe the network to have lower temporal consistencies. This helps in improving the performance of semantic segmentation models without the need of expensive labels.

4.5. Ablation Study

We perform a controlled study to isolate the effects of various optical flow techniques on the evaluation of the metric. The accuracy of estimating flow vectors and the following warping may introduce errors in the calculation of true temporal consistency (3). The accuracy of the mapping of the optical flow is calculated by measuring the MSE between the image $\mathbf{x}_t$ and the warped image $\hat{\mathbf{x}}_t$. For this purpose, varying dense optical flow methods are used. We investigate the large displacement optical flow (LDOF), that considers forward and backward flows, as described by Sundaram *et al*. [31], the flow methods introduced by Gunnar *et al*. [14], Perez *et al*. [36], and the neural network-based approach as defined by Ilg *et al*. [22]. For Ilg *et al*. [22], we do not perform any kind of fine-tuning on our datasets. The results of these experiments are shown in Table 3. We study how the mTC method is affected by different estimators of optical flow between consecutive images. An analysis of the computation time taken for the flow estimations is also performed, giving us insights into the real-time applicability of this metric. For Cityscapes, we observe in Table 3 that all the methods have very low differences in MSE and the changes in the mTC values are also very small. The error in the calculation of temporal consistency (3) due to flow vector estimation and warping are therefore small. The optical flow method from Sundaram *et al*. [31] is more accurate in mapping flow vectors and this can be seen from the MSE values, for both Cityscapes and CamVid datasets, in Table 3. The neural network-based approach defined by Ilg *et al*. [22] is the fastest but doubles the MSE values. In this paper, we adopted the approach from [31], as this has the most accurate flow vector mapping. We also performed correlation experiments with the other optical flow methods, showing similar positive strong correlations to mIoU_t. Our metric is, therefore, fairly independent of the actual method of optical flow estimation.

The neural network-based approach is the fastest in terms of flow vector calculation and an improvement in run-time and accuracy will further facilitate our metric for real-time implementations.

5. Conclusions

In this paper, we have introduced a novel mean temporal consistency (mTC) metric to measure temporal stability of the predictions of semantic segmentation models. Metrics to evaluate semantic segmentation such as mean intersection over union (mIoU) do not incorporate temporal characteristics and thus cannot be easily extended towards evaluating the consistency of the prediction of networks on video sequences. We performed an in-depth analysis of the proposed metric, study correlations to the supervised mIoU metric and find strong correlations between the two. Due to these strong correlations, our new metric can serve to select additional training data to be labelled to improve the quality of semantic segmentation networks. We performed this evaluation of temporal consistency for five state-of-the-art semantic segmentation networks of varying complexities and architectures, and on two datasets, Cityscapes and CamVid. Due to the online capability and the strong correlation to mIoU, our novel TC metric could also be used as an observer, in parallel to primary perception modules in the vehicle. Although we have performed these detailed experiments on a semantic segmentation task, the intuitive nature of our metric allows for extension to both 2D and 3D object detection tasks.

Acknowledgment

The authors would like to thank the contribution by Jonas Löhdefink and Marvin Klingner from Technische

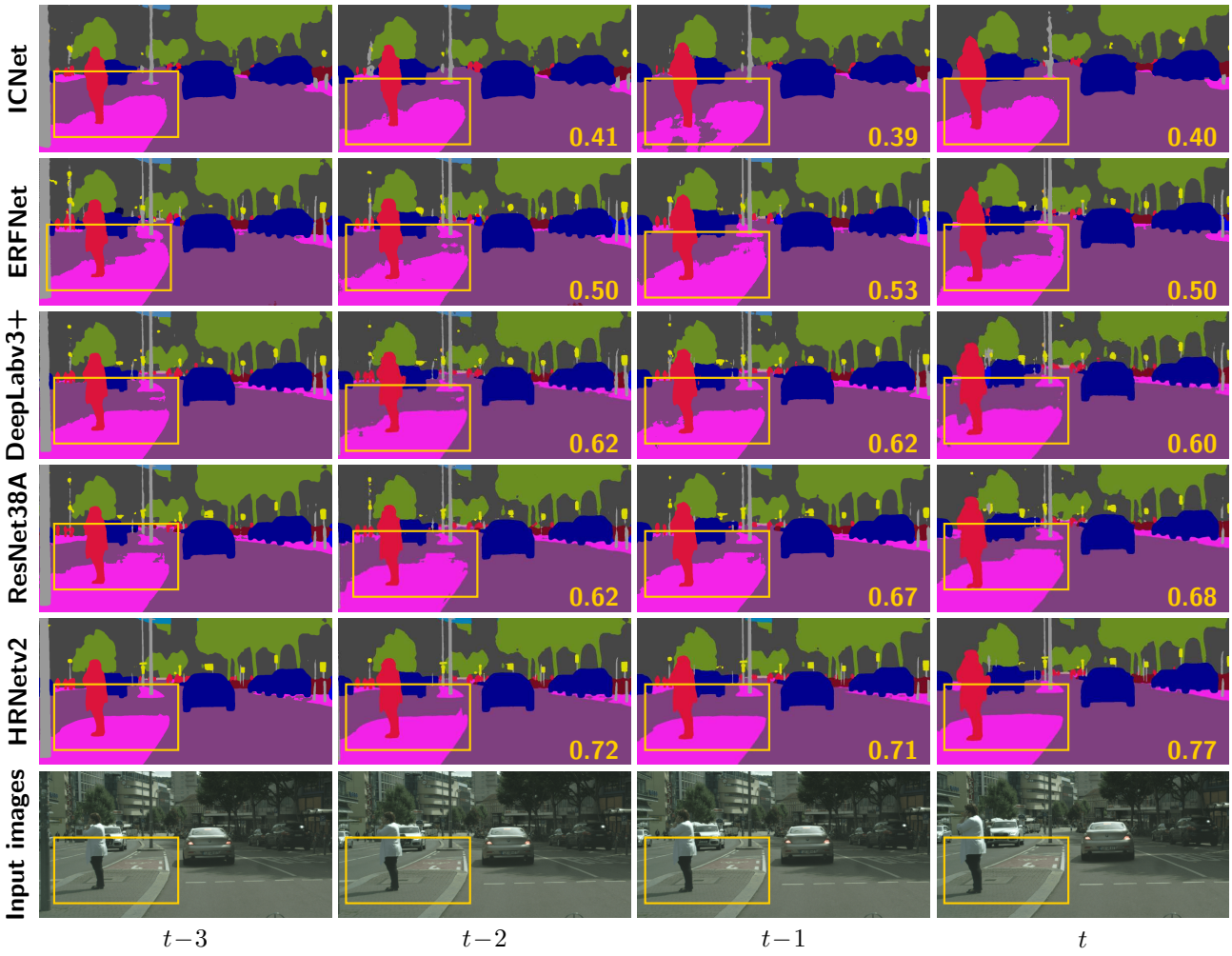

Figure 4: **Example segmentations on the Cityscapes dataset.** We show a snippet from `sequence_00`, where the instability of the networks becomes visible. The yellow boxes highlight regions of the image, where the smaller models tend to be more unstable compared to the bigger models. The numbers indicate the temporal consistency TC_τ, calculated between frame τ and $\tau-1$. As the C26 metric [26] is a percentage of consistent tracks, its values are the same as reported in Table 2. We observe the powerful HRNetv2 to be the most stable model, as also indicated by our novel mTC metric in Table 2.

Universität Braunschweig for the discussions. We also thank our colleagues of Architecture and AI Technologies, Automated Driving from Volkswagen Group Automation for the experiments and discussions.

The research leading to the results presented above are funded by the German Federal Ministry for Economic Affairs and Energy within the project "KI-Absicherung - Safe AI for automated driving".

References

[1] J. Abel, M. Kaniewska, C. Guillaumé, W. Tirry, and T. Fingscheidt. An Instrumental Quality Measure for Artificially Bandwidth-Extended Speech Signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(2):384–396, Feb. 2017. 6

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. In *Proc. of PAMI*, pp. 2481–2495, Kharagpur, India, Oct. 2016. 2

[3] P. Bilinski and V. Prisacariu. Dense Decoder Shortcut Connections for Single-Pass Semantic Segmentation. In *Proc. of CVPR*, pp. 6596–6605, Salt Lake City, UT, USA, June 2018. 2

[4] J.-A. Bolte, A. Bär, D. Lipinski, and T. Fingscheidt. Towards Corner Case Detection for Autonomous Driving. In *Proc. of IV*, pp. 438–445, Paris, France, June 2019. 2

[5] J.-A. Bolte, M. Kamp, A. Breuer, S. Homoceanu, P. Schlicht, F. Hüger, D. Lipinksi, and T. Fingscheidt. Unsupervised

Domain Adaptation to Improve Image Segmentation Quality Both in the Source and Target Domain. In *Proc. of CVPR - Workshops*, pp. 1–10, Long Beach, CA, USA, June 2019. 2, 5, 7

[6] A. Bär, F. Hüger, P. Schlicht, and T. Fingscheidt. On the Robustness of Teacher-Student Frameworks for Semantic Segmentation. In *Proc. of CVPR - Workshops*, pp. 1–9, Long Beach, CA, USA, June 2019. 2

[7] S. Caelles, A. Montes, K.-K. Maninis, Y. Chen, L. Van Gool, F. Perazzi, and J. Pont-Tuset. The 2018 DAVIS Challenge on Video Object Segmentation. *arXiv*, Mar. 2018. 3

[8] S. Caelles, J. Pont-Tuset, F. Perazzi, A. Montes, K.-K. Maninis, and L. Van Gool. The 2019 DAVIS Challenge on VOS: Unsupervised Multi-Object Segmentation. *arXiv*, May 2019. 3

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. 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, Apr. 2018. 5, 7

[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of CVPR*, pp. 3213–3223, Las Vegas, NV, USA, June 2016. 5

[11] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox. FlowNet: Learning Optical Flow with Convolutional Networks. In *Proc. of ICCV*, pp. 2758–2766, Las Condes, Chile, Dec. 2015. 3

[12] M. Everingham, S. M. Eslami, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, Jan. 2015. 2, 4

[13] G. Fabio, N. N. Shankar, C. T. Jiménez, B. Thomas, and S. Bernt. A Unified Video Segmentation Benchmark: Annotation, Metrics and Analysis. In *Proc. of ICCV*, pp. 3527–3534, Sydney, Australia, Dec. 2013. 3

[14] G. Farnebäck. Two-frame Motion Estimation Based on Polynomial Expansion. In *Proc. of SCIA*, pp. 363–370, Halmstad, Sweden, June 2003. 3, 7

[15] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. Generalisation in Humans and Deep Neural Networks. In *Proc. of NIPS*, pp. 7549–7561, Montréal, QC, Canada, Dec. 2018. 2

[16] R. Girshick. Fast R-CNN. In *Proc. of ICCV*, pp. 1440–1448, Las Condes, Chile, Dec. 2015. 2

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proc. of CVPR*, pp. 580–587, Columbus, OH, USA, June 2014. 2

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proc. of CVPR*, pp. 770–778, Las Vegas, NV, USA, June 2016. 2

[19] B. K. P. Horn and B. G. Schunck. Determining Optical Flow. *Artificial Intelligence*, 17(1–3):185–203, Aug. 1981. 3

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *Proc. of CVPR*, pp. 4700–4708, Honulu, HI, USA, July 2017. 2

[21] Z. Hui, F. Jason, and G. Sally. Image Segmentation Evaluation: A Survey of Unsupervised Methods. *Computer Vision and Image Understanding*, 110(4):260–280, Apr. 2008. 2

[22] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of Optical Flow Estimation With Deep Networks. In *Proc. of CVPR*, pp. 2704–2713, Honulu, HI, USA, July 2017. 3, 7

[23] M. Jaderberg, K. Simonyan, and A. Z. K. Kayukcuoglu. Spatial Transformer Networks. In *Proc. of NIPS*, pp. 2017–2025, Montréal, QC, Canada, Dec. 2015. 4

[24] D. Jayaraman and K. Grauman. Slow and Steady Feature Analysis: Higher Order Temporal Coherence in Video. In *Proc. of CVPR*, pp. 3852–3861, Boston, MA, USA, June 2015. 2

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. of NIPS*, pp. 1097–1105, Lake Tahoe, NV, USA, Dec. 2012. 2

[26] A. Kundu, V. Vineet, and V. Koltun. Feature Space Optimization for Semantic Video Segmentation. In *Proc. of CVPR*, pp. 3168–3175, Las Vegas, NV, USA, June 2016. 3, 5, 6, 8

[27] J. Löhdefink, A. Bär, N. M. Schmidt, F. Hüger, P. Schlicht, and T. Fingscheidt. On Low-Bitrate Image Compression for Distributed Automotive Perception: Higher Peak SNR Does Not Mean Better Semantic Segmentation. In *Proc. of IV*, pp. 352–359, Paris, France, June 2019. 5

[28] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *Proc. of CVPR*, pp. 3431–3440, Boston, MA, USA, June 2015. 2

[29] K. Maag, M. Rottmann, and H. Gottschalk. Time-Dynamic Estimates of the Reliability of Deep Semantic Segmentation Networks. *arXiv*, Nov. 2019. 3

[30] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *Proc. of ICCV*, pp. 416–423, Vancouver, Canada, July 2001. 2

[31] S. Narayanan, B. Thomas, and K. Kurt. Dense Point Trajectories by GPU-Accelerated Large Displacement Optical Flow. In *Proc. of ECCV*, pp. 438–451, Heraklion, Greece, Sept. 2010. 3, 4, 5, 6, 7

[32] D. Nilson and C. Sminchisescu. Semantic Video Segmentation by Gated Recurrent Flow Propagation. In *Proc. of CVPR*, pp. 6819–6828, Salt Lake City, UT, USA, June 2018. 3, 6

[33] H. Noh, S. Hong, and B. Han. Learning Deconvolution Network for Semantic Segmentation. In *Proc. of ICCV*, pp. 1520–1528, Santiago, Chile, Dec. 2015. 2

[34] S. Paul, A. Karteek, L. Lubor, and T. Philip. Combining Appearance and Structure from Motion Features for Road Scene Understanding. In *Proc. of BMVC*, pp. 1–11, London, England, Sept. 2009. 6

[35] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *Proc. of CVPR*, pp. 724–732, Las Vegas, NV, USA, June 2016. 3

[36] J. S. Pérez, E. Meinhardt, and G. Facciolo. TV-L1 Optical Flow Estimation. *IPOL Journal*, 3:137–150, Apr. 2013. 3, 7

[37] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv*, Mar. 2017. 3

[38] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *Proc. of CVPR*, pp. 779–788, Las Vegas, NV, USA, June 2016. 2

[39] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection With Region Proposal Networks. In *Proc. of NIPS*, pp. 91–99, Montréal, QC, Canada, Dec. 2015. 2

[40] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo. ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, Jan. 2018. 5, 7

[41] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proc. of MICCAI*, pp. 234–241, Munich, Germany, Oct. 2015. 2

[42] T. Sämann, K. Amende, S. Milz, and H.-M. Groundefined. Robust Semantic Video Segmentation Through Confidence-Based Feature Map Warping. In *Proc. of CSCS*, pp. 1–9, Kaiserslautern, Germany, Oct. 2019. 2

[43] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. of ICLR*, pp. 1–27, San Diego, CA, USA, May 2015. 2

[44] S. Sun, J. Pang, J. Shi, S. Yi, and W. Ouyang. FishNet: A Versatile Backbone for Image, Region, and Pixel Level Prediction. In *Proc. of NIPS*, pp. 754–764, Montréal, QC, Canada, Dec. 2018. 2

[45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. In *Proc. of CVPR*, pp. 1–9, Boston, MA, USA, June 2015. 2

[46] S. Ulbrich, T. Menzel, A. Reschka, F. Schuldt, and M. Maurer. Defining and Substantiating the Terms Scene, Situation, and Scenario for Automated Driving. In *Proc. of ITSC*, pp. 982–988, Canary Islands, Spain, Sept. 2015. 2

[47] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep High-Resolution Representation Learning for Visual Recognition. *arXiv*, Aug. 2019. 5, 6, 7

[48] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In *Proc. of ECCV*, pp. 405–420, Munich, Germany, Sept. 2018. 5, 7