Perceptual Consistency in Video Segmentation

Yizhe Zhang [*†]
yizhe.zhang.cs@gmail.com

Shubhankar Borse [‡§]
sborse@qti.qualcomm.com

Hong Cai [‡§]
hongcai@qti.qualcomm.com

Ying Wang [*]
yingwang0022@gmail.com

Ning Bi [¶]
nbi@qti.qualcomm.com

Xiaoyun Jiang [¶]
xjiang@qti.qualcomm.com

Fatih Porikli [‡]
fporikli@qti.qualcomm.com

Abstract

*In this paper, we present a novel **perceptual consistency** perspective on video semantic segmentation, which can capture both temporal consistency and pixel-wise correctness. Given two nearby video frames, perceptual consistency measures how much the segmentation decisions agree with the pixel correspondences obtained via matching general perceptual features. More specifically, for each pixel in one frame, we find the most perceptually correlated pixel in the other frame. Our intuition is that such a pair of pixels are highly likely to belong to the same class. Next, we assess how much the segmentation agrees with such perceptual correspondences, based on which we derive the perceptual consistency of the segmentation maps across these two frames. Utilizing perceptual consistency, we can evaluate the temporal consistency of video segmentation by measuring the perceptual consistency over consecutive pairs of segmentation maps in a video. Furthermore, given a sparsely labeled test video, perceptual consistency can be utilized to aid with predicting the pixel-wise correctness of the segmentation on an unlabeled frame. More specifically, by measuring the perceptual consistency between the predicted segmentation and the available ground truth on a nearby frame and combining it with the segmentation confidence, we can accurately assess the classification correctness on each pixel. Our experiments show that the proposed perceptual consistency can more accurately evaluate the temporal consistency of video segmentation as compared to flow-based measures. Furthermore, it can help more confidently predict segmentation accuracy on unlabeled test frames, as compared to using classification confidence alone. Finally, our proposed measure can be used as a regularizer during the training of segmentation models, which leads to more temporally consistent video segmentation while maintaining accuracy.*

1. Introduction

High-quality video semantic segmentation is a critical task for a large variety of downstream applications, such as video processing, AR/VR, robotics, and autonomous driving. To be deployed for practical use, the segmentation model needs to be not only pixel-wise accurate, but also temporally consistent when applied to videos. In order to facilitate the development of such high-quality models, it is necessary to be able to assess both pixel-wise accuracy and temporal consistency accurately.

While segmentation accuracy can be easily evaluated given the ground-truth labels on the test data, assessing the temporal consistency is not straightforward. In earlier work, Perazzi et al. [27] measure the temporal consistency of object segmentation by computing the distance between the segmentation masks across two consecutive frames. This approach, however, does not factor in the object movements and changing occlusions. Most of the recent works [16, 24, 33] utilize motion-based pixel correspondence between two consecutive frames (i.e., optical flow [13]), to measure temporal consistency. More specifically, given two consecutive video frames, the segmentation of one frame is warped to the other based on the estimated flow, and the warped and actual segmentation maps are then compared to measure the segmentation consistency between these two frames. However, it is challenging to generate highly accurate and generalizable optical flow estimations. Furthermore, such exact pixel correspondence is susceptible to occlusions and objects moving out of the frame.

When ground-truth annotations are not available during test/deployment, we can no longer use the standard accuracy metrics (e.g., mIoU, pixel-wise accuracy). For instance, in popular video semantic segmentation datasets such as Cityscapes [3] and CamVid [1], the test video frames are only sparsely labeled due to the high annotation costs. When ground-truth labels are not available on a test frame, one can
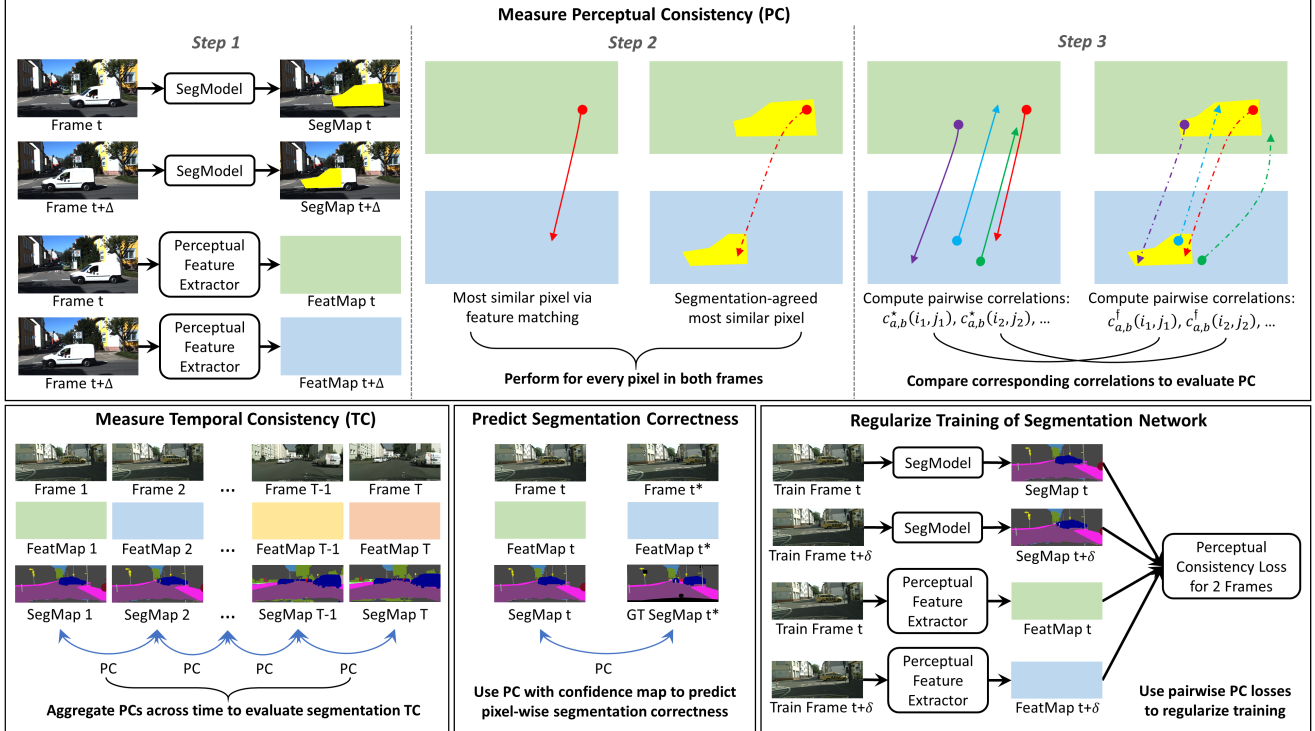
Figure 1: **Top:** Overview of measuring perceptual consistency of segmentation decisions on two nearby frames. **Bottom:** Use cases of perceptual consistency, including measuring temporal consistency of video semantic segmentation, predicting pixel-wise segmentation correctness on unlabeled test video frames, and regularizing the training of a segmentation network.

use the pixel-wise classification confidence scores to predict the uncertainty or correctness of the segmentation network. However, this only captures the network's own output on the unlabeled frame and does not utilize the sparsely labeled frames.

In this work, we present a new perspective on the quality of video semantic segmentation models, i.e., **perceptual consistency**. Our intuition is that given two nearby video frames, which share similar visual contents, pixels that are highly perceptually similar are highly likely to belong to the same class and thus, should receive the same label from the segmentation model (SegModel). In order to concretize this, given two nearby video frames and the predicted segmentation maps (SegMaps) on them, we assess how much the segmentation agrees with the cross-frame pixel correspondence established on the two frames' perceptual feature maps (FeatMaps). More specifically, for each pixel in one frame, we first find the most correlated pixel from the other frame by matching perceptual features. We consider these two pixels are expected to belong to the same class. We next find the most correlated pixel that is also agreed by the segmentation maps. If the segmentation agrees with the perceptual correspondence, the correlations found via unconstrained feature matching and segmentation-agreed feature matching will be equal; otherwise, the segmentation-constrained one will be

smaller. As such, we can use the ratio between these two correlations to quantify the (pixel-wise) agreement between the segmentation decisions and the perceptual correspondence. This ratio can then be aggregated over the pixels to measure the perceptual consistency between the segmentation maps on the two frames. This procedure is illustrated in Fig. 1 (top).

We can see that, unlike optical flow, our measure does not seek exact pixel correspondence across two images, which requires each pair of corresponding pixels to associate with the same object point; it instead finds pairs of maximally correlated pixels. This relaxation makes our measure robust to cases where exact correspondence does not exist, e.g., due to occlusions.

By measuring the perceptual consistency between segmentation maps on consecutive pairs of frames, we can naturally capture the temporal consistency of video segmentation. In addition, in the case of a sparsely labeled test video, we can utilize perceptual consistency (between an unlabeled frame and a nearby labeled frame), in addition to the segmentation confidence, to accurately predict the pixel-wise segmentation correctness on an unlabeled test frame. This allows the prediction to leverage information beyond a single frame. Furthermore, we can integrate perceptual consistency as an additional regularization in training (which does not

incur extra computation during inference). This enables the trained video segmentation network to generate temporally more consistent results while maintaining accuracy. These use cases of perceptual consistency are illustrated in Fig. 1 (bottom).

We summarize our main contributions as follows:

- We propose a novel perceptual consistency measure that evaluates how much the segmented predictions over two similar images agree with the pixel correspondence found via matching general perceptual features.

- Our perceptual consistency measure, when applied to segmented predictions on consecutive pairs of frames in a video, can more accurately evaluate the temporal consistency, as compared to the existing flow-based measure.

- Given a non-training image that does not have the ground-truth annotation, our measure can be used to accurately predict the pixel-wise correctness of the segmentation on this image by cross-referencing the available ground-truth of a similar image.

- Furthermore, our proposed measure can be used as a regularizer during training, which leads to more temporally consistent segmentation models while maintaining accuracy; this property holds even when compared to models trained with optical flow.

2. Related Work

Video Semantic Segmentation: Video segmentation has been a major goal in computer vision research. In addition to improving the accuracy of semantic predictions [6, 31, 32], previous works have studied various other aspects that are more specific to the video setting, e.g., efficient processing [11, 18, 23], label propagation [25, 35], and temporal consistency [12, 19, 24]. We refer readers to recent surveys [21, 36] for a more comprehensive discussion.

Temporal Consistency: Since accuracy measures (e.g., mIoU) cannot properly capture temporal consistency, researchers have proposed various ways to measure segmentation temporal consistency. For instance, [27] computes the distance between the segmentation masks from two consecutive frames. This, however, does not factor in motion. As such, researchers incorporate motion estimation (e.g., optical flow) when measuring temporal consistency [16, 24, 19, 33]. However, estimating accurate flow on real-world data can be very challenging, and in many cases, more error-prone and time-consuming than the segmentation task itself.

In order to improve temporal consistency, previous works have designed various models to take in additional information, e.g., optical flow [12, 24], 3D structure [7, 15], and temporal correlation [11, 16, 28, 30]. However, these methods require multi-frame information to segment each frame, incurring extra computation overhead. Recently, Liu et al. [19] have proposed using optical flow only during training and performing per-frame inference. However, their method is still constrained by the accuracy of the estimated flow.

Predicting Segmentation Correctness: Researchers have studied various ways to capture the output uncertainty/correctness of a neural network for classification tasks, e.g., maximum softmax probability (i.e., confidence score) [10, 8], dropout/sampling [14], using ensembles [17], stochastic variational Bayesian inference [26]. However, during test time, these methods only utilize the information from the single unlabeled input image and can suffer from domain shift [26]. Moreover, except for the case of using classification confidence, the existing methods require specific training schemes and/or network modules. As such, they cannot be applied to any given trained model.

3. Perceptual Consistency

Perceptual consistency is our novel take on assessing the quality of video semantic segmentation. Given two nearby video frames, perceptual consistency measures how much the segmentation agrees with the cross-frame perceptual pixel correspondence. Perceptual consistency can be applied to the segmentation maps on consecutive pairs of frames in a video, in order to evaluate the segmentation temporal consistency. Furthermore, given a sparsely labeled test video, perceptual consistency can facilitate the prediction of pixel-wise correctness of the segmentation on a frame that does not have ground-truth annotation. Finally, we can utilize perceptual consistency to regularize the training of a segmentation network.

3.1. Measuring Perceptual Consistency

Consider two nearby video frames which share similar visual contents, x_a and x_b, and their respective segmentation maps (e.g., predicted by a neural network), y_a and y_b. Let f_a and f_b denote the general perceptual feature maps for x_a and x_b, which can be obtained from an image representation network trained on a large, general dataset (e.g., a ResNet [9] trained on ImageNet [4]).

Perceptual Correspondence: For every pixel (i, j) in x_a, we find its perceptually corresponding pixel in x_b by solving the following maximization:

$$c_{a,b}^\star(i,j; f_a, f_b) = \max_{i',j'} \sigma(f_a(i,j), f_b(i',j')), \quad (1)$$

where σ computes the cosine similarity between two feature vectors and $c_{a,b}^\star(i,j; f_a, f_b)$ is the maximum correlation between pixel (i, j) in x_a and the pixels in x_b, given the features maps, f_a and f_b. We denote the most correlated pixel in x_b as $(i^\star, j^\star)$.

The pixel $(i^\star, j^\star)$ in x_b is the most perceptually similar to pixel (i, j) in x_a. Since x_a and x_b are neighboring frames

and share highly overlapping visual contents, it is highly likely that these two pixels belong to the class. For instance, on MIT DriveSeg [5], a pair of perceptually corresponding pixels from two consecutive frames have a probability of 0.944 to be in the same class.

Segmentation-Agreed Correspondence: Next, we find the pixel correspondence that is also agreed by the segmentation maps. More specifically, for pixel (i, j) in x_a, we solve for the most perceptually correlated pixel in x_b with the additional constraint that the two pixels are labeled with the same class on the segmentation maps:

$$c_{a,b}^{\dagger}(i, j; f_a, f_b, y_a, y_b) =$$
$$\max_{i',j'} \sigma(f_a(i,j), f_b(i',j')), \text{ s.t. } y_a(i,j) = y_b(i',j'), \quad (2)$$

where $c_{a,b}^{\dagger}(i, j; f_a, f_b, y_a, y_b)$ is the segmentation-agreed maximum correlation between pixel (i, j) in x_a and the pixels in x_b, given the features maps, f_a and f_b, and the segmentation maps, y_a and y_b. we denote the found pixel in x_b as $(i^{\dagger}, j^{\dagger})$.

Measuring Consistency: If $i^{\star} = i^{\dagger}$ and $j^{\star} = j^{\dagger}$, then the segmentation decisions completely agree with the perceptual correspondence that these two pixels should belong to the same class. As a result, we have $c_{a,b}^{\star}(i, j) = c_{a,b}^{\dagger}(i, j)$.[1] Otherwise, $c_{a,b}^{\dagger}(i, j)$ will be smaller and its value depends on how much the segmentation decisions align with the perceptual correspondence. More specifically, if $x_b(i^{\dagger}, j^{\dagger})$ is still highly perceptually correlated to $x_a(i, j)$ (while not being the most correlated one), then $c_{a,b}^{\dagger}$ will still be close to $c_{a,b}^{\star}(i, j)$. This indicates that the segmentation decisions highly agree with the perceptual correspondence. On the other hand, if $x_b(i^{\dagger}, j^{\dagger})$ is not perceptually similar to $x_a(i, j)$, then $c_{a,b}^{\dagger}(i, j)$ will be much smaller than $c_{a,b}^{\star}(i, j)$. In this case, the segmentation decisions disagree with the perceptual correspondence, and do not assign the same class to $x_a(i, j)$ and its perceptually similar pixels in x_b. As such, by comparing $c_{a,b}^{\dagger}(i, j)$ and $c_{a,b}^{\star}(i, j)$, we can capture the pixel-wise consistency between the segmentation decisions and the perceptual correspondence. We refer to such a consistency as the **perceptual consistency**.

To quantify the perceptual consistency of the segmentation decisions over the two frames, we compute the ratio of $c_{a,b}^{\dagger}(i, j)$ and $c_{a,b}^{\star}(i, j)$, and then aggregate it over the frames as follows:
$$\rho(y_a, y_b) =$$
$$\frac{1}{H \times W} \min \left\{ \sum_{i,j} \frac{c_{a,b}^{\dagger}(i,j) - \bar{c}_{a,b}^{\dagger}}{c_{a,b}^{\star}(i,j) - \bar{c}_{a,b}^{\star}}, \sum_{i,j} \frac{c_{b,a}^{\dagger}(i,j) - \bar{c}_{b,a}^{\dagger}}{c_{b,a}^{\star}(i,j) - \bar{c}_{b,a}^{\star}} \right\}, \quad (3)$$

---
[1] We omit the given parameters listed after the semicolon when the context is clear.

where $\bar{c}_{a,b}^{\dagger}$, $\bar{c}_{b,a}^{\star}$, $\bar{c}_{a,b}^{\dagger}$, and $\bar{c}_{b,a}^{\star}$ are the respective frame-level averages, and H and W are the frame height and width. We take the minimum of the two frame-level perceptual consistencies to better capture the disagreements between segmentation decisions and perceptual correspondences.

3.2. Perceptual Consistency for Measuring Video Segmentation Temporal Consistency

In a video (of length T), two consecutive frames share highly overlapping visual contents. As such, the perceptual consistency of Eq. 3 can be readily applied to the segmentation maps of each consecutive pair of video frames. By aggregating the perceptual consistency over all the consecutive frame pairs, we obtain a measure of the temporal consistency of the video segmentation:

$$\tilde{\rho}(y_1, y_2, ..., y_T) = \frac{1}{T-1} \sum_{t=1}^{T-1} \rho(y_t, y_{t+1}). \quad (4)$$

Robustness of Perceptual Consistency: Perceptual consistency offers a more robust way to measure temporal consistency as compared to optical flow, which most existing works use. For instance, when there is occlusion, optical flow will fail to find correspondence and thus cannot properly measure consistency for the affected image regions. On the other hand, perceptual consistency does not require each pair of perceptually corresponding pixels to associate with the same object point. Instead, we look for a pair of pixels that are the most perceptually correlated, which can always be found.

In addition, flow estimation errors can cause underestimation of the segmentation temporal consistency. For instance, when the estimated flow incorrectly connects two pixels that do not belong to the same class, the flow-based measure will assess zero consistency for this pair even when the segmentation decisions are correct. On the other hand, while it can occasionally happen that two perceptually corresponding pixels from two consecutive frames (found via Eq. 1) do not belong to the same class (e.g., with a probability of 0.056 on MIT DriveSeg), the probability is very high that given a pixel in one frame, there is a same-class pixel among the highly perceptually correlated pixels in the other frame. For instance, on MIT DriveSeg, this probability is 0.989 if we consider the top-10 most correlated pixels, and the top-1 and top-10 correlation values differ only by 0.026. As such, the segmentation model can still achieve a reasonably high consistency score as long as it correctly labels the perceptually similar pixels.

3.3. Perceptual Consistency for Predicting Pixel-Wise Segmentation Correctness

In practice, the ground-truth labels may not be fully available during test. In such cases, we can predict the pixel-wise segmentation correctness on an unlabeled test frame

by jointly using the segmentation confidence map (model-internal image-level measure) and the perceptual consistency map between the predicted segmentation on the test frame and the available ground truth of a nearby frame (model-external temporal-level measure).

Consider a sequence of test video frames, $X = \{x_1, x_2, ..., x_T\}$, and the predicted segmentation maps, $Y = \{y_1, y_2, ..., y_T\}$. The ground-truth segmentation maps are only available for a small subset of the frames, $S = \{s_t \mid t \in \Omega_L\}$, where Ω_L is the set of time instances where ground truths are available. This setup captures the structure of common video semantic segmentation datasets, such as Cityscapes [3] and CamVid [1]

Given an unlabeled test frame, x_{t_U}, we can evaluate the quality of its segmentation by using both the decision confidence and perceptual consistency. For each pixel (i, j), the confidence is given by $z_{t_U}(i, j) = \max_{k \in \Omega_K} \tilde{z}_{t_U}(i, j, k)$, where $\tilde{z}_{t_U}(i, j, k)$ denotes the classification score for each class $k \in \Omega_K$ (after softmax). We then measure the pixel-wise perceptual consistency of y_{t_U} w.r.t. the closest available ground-truth segmentation, s_{t_L}, where $t_L = \min_{\tau \in \Omega_L} \|\tau - t_U\|$, as follows:

$$\hat{\rho}_{t_U}(i, j) = \frac{c^{\dagger}_{t_U, t_L}(i, j; \ f_{t_U}, f_{t_L}) - \bar{c}^{\dagger}_{t_U, t_L}}{c^{\star}_{t_U, t_L}(i, j; \ f_{t_U}, f_{f_L}, y_{t_U}, s_{t_L}) - \bar{c}^{\star}_{t_U, t_L}}, \quad (5)$$

where the perceptual consistency is computed only in one direction from x_{t_U} to x_{t_L}. This is because we only need to assess the consistency of the pixel-wise classification decisions on the unlabeled test frame w.r.t. the ground truth in this case.

The final pixel-wise segmentation correctness is predicted by

$$\alpha_{t_U}(i, j) = z_{t_U}(i, j) + \hat{\rho}_{t_U}(i, j), \quad (6)$$

where perceptual consistency term allows the prediction to cross-reference the available ground truth of a nearby frame, in addition to using the network's confidence on the input unlabeled frame.

3.4. Perceptual Consistency for Regularizing Network Training

We can utilize perceptual consistency (which is differentiable) as an additional regularization when training any video segmentation network. This allows the network to learn to generate more perceptually consistent segmentation results across consecutive frames, which can improve temporal consistency. This is done without any change to the network. As such, our method provides an improvement without incurring extra computational cost during inference.

Given N training videos, the perceptual consistency loss is given as follows:

$$\mathcal{L}_{PC} = \frac{1}{N} \sum_{v=1}^{N} \frac{1}{T_v \times T_w} \sum_{i=1}^{T_v} \sum_{\delta=1}^{T_w} 1 - \rho(x_t^v, x_{t+\delta}^v), \quad (7)$$

where $\rho(x_t^v, x_{t+\delta}^v)$ is computed using Eq. 3 for frames x_t^v and $x_{t+\delta}^v$ from the v-th video, T_v is the length of video v, and T_w is a time window within which we compute the loss for all the frame pairs.

The perceptual consistency loss of Eq. 7 is then combined with the standard multi-class cross-entropy loss for segmentation [22]. The total loss for training the network is then given by

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{TC}}, \quad (8)$$

where λ balances the cross-entropy loss and the perceptual consistency loss.

4. Experiments

In this section, we conduct extensive experiments to demonstrate the efficacy of our proposed perceptual consistency, in terms of measuring the temporal consistency of video semantic segmentation, aiding the prediction of pixel-wise segmentation correctness, and regularizing the training of a segmentation network to promote better temporal consistency. We further conduct ablation studies to analyze the choice of perceptual features.

4.1. Experiment Setup

Datasets: For evaluating temporal consistency measures, we use MIT DriveSeg [5] and DAVIS 2016 [27] where per-frame ground-truth segmentation is available, and KITTI Flow 2015 [20] where ground-truth optical flow is available. DriveSeg contains a video sequence of 5,000 frames with 15 semantic classes. The first 4,500 frames are used for training and the remaining 500 frames for test. DAVIS contains a total of 50 video sequences, with 3455 annotated frames. It is designed for foreground object segmentation. To measure TC, we use the validation set, consisting of 20 videos. In each frame, the ground-truth segmentation is provided to separate the object of interest from the background. For evaluating the prediction of segmentation correctness, we use DriveSeg. To evaluate the usage of perceptual consistency as a training regularization, we use Cityscapes [3] and CamVid [1], which are standard video semantic segmentation benchmarks.

Perceptual Features: In our main experiments, we use a ResNet-18 [9] trained on ImageNet [4] to extract perceptual features. We also include analysis of other feature extraction networks in the ablation studies, such as ResNet-101 [9] and WideResNet-50 (WRN-50) [37].

Segmentation Networks: We use state-of-the-art segmentation models in our experiments. For evaluating TC, on DriveSeg, we use HRNet-w48 [34] and PSPNet [38] with ResNet-101 backbone that are trained on DriveSeg. On DAVIS, we use the trained models of STM [25] and RANet [35] from the official repositories. On KITTI Flow, we use two HRNet-w48 models, one trained on DriveSeg and

the other trained on Cityscapes. For evaluating the prediction of segmentation correctness, we use the PSPNet trained on DriveSeg. For the training experiments on Cityscapes and CamVid, we use HRNet-w18 [34] and DeepLabV3+ [2] with ResNet-101 backbone.

Temporal Consistency Measure: We consider the flow-based TC as the ground-truth TC when using ground-truth flows, which computes the mIoU between warped and actual segmentation maps across frames. For a practical baseline, we use the flow-based measure in [19] which uses a pretrained FlowNetV2 [29] to estimate optical flow.

4.2. Measuring Temporal Consistency

To evaluate the ability of perceptual consistency to measure segmentation temporal consistency (TC), we conduct experiments on the videos from DriveSeg and DAVIS. We further evaluate on KITTI Flow where we utilize the ground-truth optical flows to compute ground-truth TC.

Measuring TC for Ground-Truth Segmentation: We first perform a verification experiment where we measure the temporal consistency of ground-truth segmentation, for which an accurate measure should return a TC value close to 1. As shown in Table 1, perceptual consistency assesses a close-to-1 TC for the ground-truth video segmentation. On the other hand, the flow-based measure assesses much lower TC values for the ground truth.

Evaluation Based on Per-Frame Ground-Truth Segmentation: For a video, given the predicted segmentation, $y_1, y_2, ..., y_T$, and the ground-truth segmentation, $s_1, s_2, ..., s_T$, we can construct a sequence of alternating predicted and GT segmentation maps (e.g., $y_1, s_2, y_3, s_4, ...$), for which the ground-truth TC can be derived. In this sequence, for a pair, y_t and s_{t+1}, we can use s_t to approximate the warped version of s_{t+1} at t under the assumption that one-to-one pixel correspondence exists between the two frames. The mIoU between y_t and s_t then captures the ground-truth TC between y_t and s_{t+1}. We evaluate the correlation between perceptual consistency and the ground-truth TC for such alternating segmentation sequences. As shown in Tables 2 and 3, our perceptual consistency correlates much better with the ground-truth TC, in terms of the Pearson, Spearman, and Kendall correlation coefficients, as compared to the flow-based measure.

Evaluation Based on Ground-Truth Flow: We can obtain the ground-truth TC on KITTI Flow, by using the available ground-truth flow to warp the segmentation map from one frame to another and compute the mIoU between the warped and actual maps for every consecutive pair of frames. As shown in Table 4, in most cases, our perceptual consistency correlates significantly better with the ground-truth TC as compared to the measure using estimation flows.

4.3. Using Perceptual Consistency in Training

On Cityscapes and CamVid, we take a pretrained segmentation network (using only cross-entropy) and use the total loss of Eq. 8 (with $\lambda = 0.5$) to finetune it for 20 epochs. We use stochastic gradient descent, and set the momentum to 0.9, weight decay to 0.0005, and learning rate to 0.0001.

It can be seen in Table 5 that by using our perceptual consistency regularization, we can improve the temporal consistency of the segmentation network while slightly improving or maintaining the accuracy. As compared to using a flow-based regularization [19], perceptual consistency is more effective in promoting the model's temporal consistency. Note that we use the flow-based measure (with FlowNetV2) in [19] to evaluate temporal consistency, in order to compare with reported results in the literature (although the estimated flows can contain errors).

4.4. Predicting Pixel-Wise Segmentation Correctness

We use Eq. 6 to predict the pixel-wise correctness of the segmentation, which fuses the classification confidence map and the perceptual consistency map of Eq. 5. In Fig. 2, we show the Receiver Operating Characteristic (ROC) and precision-recall (PR) curves of our correctness prediction, for different sparsity levels of the available ground-truth segmentation maps. It can be seen that it is not sufficient to the confidence alone for the prediction (blue), since it only utilizes the information from a single unlabeled test image. Adding perceptual consistency considerably improves the prediction by exploiting the information from a nearby labeled frame (green). For instance, when a ground-truth segmentation map is available every 4 frames, adding perceptual consistency improves the prediction AUC from 0.897 to 0.938. This means that even if only 25% of the video frames are annotated, our approach can still accurately predict the pixel-wise segmentation correctness for the unlabeled frames. It is noteworthy that confidence and perceptual consistency capture different characteristics of the correctness, as can be seen in the figure. As such, fusing them enhances the overall prediction.

Since optical flow also captures inter-frame information, we experiment with combining confidence and the flow-based measure to predict correctness (purple). This prediction is less accurate as compared to the case of adding perceptual consistency. Moreover, since the flow-based measure returns only a binary result for each pixel, it has only one point in the ROC plot. Such binary inputs introduce irregularity into the prediction, as shown by the less regular shape of the purple ROC curve.

4.5. Additional Study

In this part, we study the effect of using different perceptual feature extraction networks. We repeat the experiments

Method	MIT DriveSeg	DAVIS
Flow-Based (FlowNetV2)	0.8682 ± 0.0633	0.9454 ± 0.0540
Perceptual Consistency	0.9847 ± 0.0132	0.9985 ± 0.0014

Table 1: Measuring TC of ground-truth segmentation on DriveSeg and DAVIS. We report the average TC values across time, as well as the standard deviations.

Method	Pearson		Spearman		Kendall	
	HRNet-w48	PSPNet	HRNet-w48	PSPNet	HRNet-w48	PSPNet
Flow-Based (FlowNetV2)	0.5768	0.9459	0.5568	0.8568	0.4212	0.7118
Perceptual Consistency	0.7432	0.9492	0.8170	0.8602	0.6599	0.7293

Table 2: Correlations between measured TC and ground-truth TC on DriveSeg. A higher correlation indicates a more accurately measured TC. P-values for all reported correlations are smaller than 0.0001.

Method	Pearson		Spearman		Kendall	
	STM	RANet	STM	RANet	STM	RANet
Flow-Based (FlowNetV2)	0.6627	0.6984	0.7030	0.7131	0.5218	0.5263
Perceptual Consistency	0.8048	0.8551	0.7297	0.8304	0.5393	0.6366

Table 3: Correlations between measured TC and ground-truth TC on DAVIS. A higher correlation indicates a more accurately measured TC. P-values for all reported correlations are smaller than 0.0001.

Method	Pearson		Spearman		Kendall	
	HRN (C)	HRN (D)	HRN (C)	HRN (D)	HRN (C)	HRN (D)
Flow-Based (FlowNetV2)	0.7124	0.7857	0.7654	0.7605	0.6028	0.5970
Perceptual Consistency	0.6808	0.8566	0.8022	0.8833	0.6113	0.7183

Table 4: Correlations between measured TC and ground-truth TC computed from ground-truth flows on KITTI Flow 2015. A higher correlation indicates a more accurately measured TC. P-values for all reported correlations are smaller than 0.0001. HRN (C) and HRN (D) denote the HRNet-w48 models trained on Cityscapes and DriveSeg, respectively.

of Table 2 and Table 4 using different ImageNet-trained networks to extract feature maps for computing perceptual consistency. We use a DriveSeg-trained HRNet-w48 as the segmentation model. It can be seen in Table 6 that when using different perceptual features, the TC measured by perceptual consistency has significantly higher correlations with the ground-truth TC, as compared to the flow-based measure.

5. Discussion

Above, we demonstrated our Perceptual Consistency (PC) is robust and accurate in measuring consistency and accuracy for video segmentation. Here, we discuss limitations and potential future work. PC requires finding dense matching for the perceptual features along the two directions, which at this moment is expected to be performed on a modern GPU/TPU for a decent throughput (e.g., under 100 ms per query). To demonstrate the robustness and generalization of PC, no special designed local window is used for the matching part in PC. However, if additional prior knowledge is available of the videos in hand, one may consider designing a way to dynamically change the search window for the matching step to save time and computation. To further

improve the robustness of PC, one can use multiple representation networks to derive more robust and general perceptual features. A new self-supervised learning task might be possible for learning better feature representations for measuring perceptual consistency and beyond. In addition, it may be possible to cross-reference (with proper weighting) multiple GT-available frames to enhance prediction accuracy for predicting segmentation correctness.

6. Conclusion

This paper proposed a novel perceptual consistency measure to capture inconsistency and inaccuracy in video segmentation. Our framework draws its quality measure via cross-referencing segmentation among similar-looking images and builds a matching paradigm that can effectively and accurately detect and measure inconsistency and inaccuracy in segmentation. Combining network confidence measure (a model-internal image-level measure) and our perceptual consistency measure (a model-external temporal-level measure), we demonstrated that it is feasible to achieve high prediction accuracy in predicting pixel-wise segmentation consistency and correctness for video segmentation.
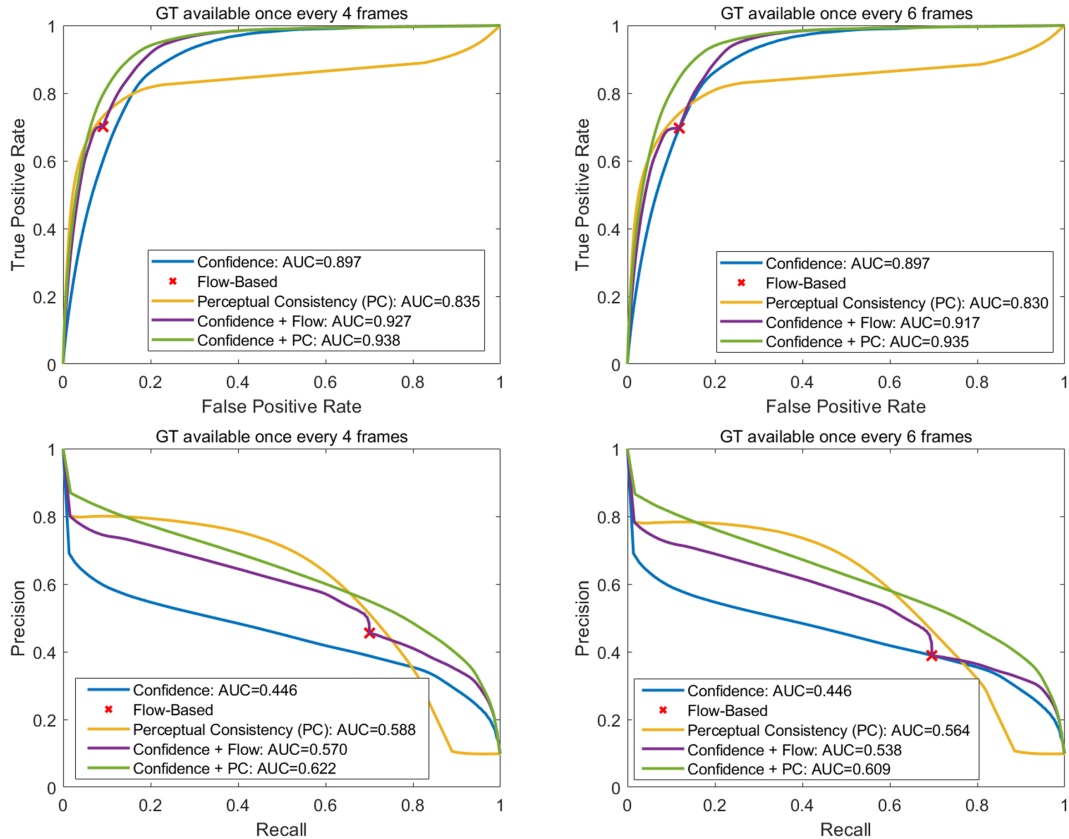
Figure 2: Receiver operating characteristic curves (top) and precision-recall curves (bottom) for predicting pixel-wise segmentation correctness on DriveSeg, for the cases where a ground-truth segmentation map is available every 4 frames (left) and every 6 frames (right). An incorrectly segmented pixel is considered as a positive sample and a correct one is a negative sample. The segmentation model is an PSPNet (with ResNet-101 backbone) trained on DriveSeg.

Model and Training Scheme	Cityscapes		CamVid	
	mIoU	Flow-Based TC	mIoU	Flow-Based TC
HRNet-w18 [34]	0.762	0.691	0.732	0.752
+ Flow-Based Loss [19]	0.764	0.696	–	–
+ Perceptual Consistency	0.764	**0.712**	0.732	**0.762**
DeepLabV3+ (ResNet-101) [2]	0.762	0.710	0.752	0.756
+ Perceptual Consistency	0.763	**0.724**	0.752	**0.765**

Table 5: Training performance of utilizing perceptual consistency as an additional regularization, as compared to the cases of no regularization and using flow-based regularization [19]. In this evaluation, we use the flow-based measure (with FlowNetV2) in [19] to assess temporal consistency in order to compare with reported results in the literature (although the estimated flows can contain errors).

Method	Pearson		Spearman		Kendall	
	DriveSeg	KITTI	DriveSeg	KITTI	DriveSeg	KITTI
Flow-Based (FlowNetV2)	0.5768	0.7857	0.5568	0.7605	0.4212	0.5970
Perceptual Consistency (ResNet-18)	0.7432	0.8566	0.8170	0.8833	0.6599	0.7183
Perceptual Consistency (ResNet-101)	0.7245	0.8127	0.8030	0.8085	0.6423	0.6273
Perceptual Consistency (WRN-50)	0.7448	0.8504	0.8188	0.8583	0.6608	0.6896

Table 6: Correlations between measured TC and ground-truth TC. We show the performance of our perceptual consistency when using different feature extractors, such as ResNet-18, ResNet-101, and WRN-50. DriveSeg denotes the evaluation setting of Table 2 and KITTI denotes the evaluation setting of Table 4.

References

[1] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *Proceedings of the European Conference on Computer Vision*, pages 44–57. Springer, 2008.

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

[3] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[5] Li Ding, Jack Terwilliger, Rini Sherony, Bryan Reimer, and Lex Fridman. MIT DriveSeg (Manual) Dataset for Dynamic Driving Scene Segmentation. Technical report, Massachusetts Institute of Technology, 2020.

[6] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10713–10720, 2020.

[7] G. Floros and B. Leibe. Joint 2D-3D temporally consistent semantic segmentation of street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[10] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of the International Conference on Learning Representations*, 2017.

[11] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8818–8827, 2020.

[12] J. Hur and S. Roth. Joint optical flow and temporally consistent semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2016.

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[14] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances on Neural Information Processing Systems*, 2017.

[15] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. M. Rehg. Joint semantic segmentation and 3D reconstruction from monocular video. In *Proceedings of the European Conference on Computer Vision*, 2014.

[16] Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun. Feature space optimization for semantic video segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3168–3175, 2016.

[17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances on Neural Information Processing Systems*, 2017.

[18] Yule Li, Jianping Shi, and Dahua Lin. Low-latency video semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5997–6005, 2018.

[19] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *Proceedings of the European Conference on Computer Vision*, pages 352–368. Springer, 2020.

[20] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, 2015.

[21] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *arXiv preprint arXiv:2001.05566*, 2020.

[22] Kevin P Murphy. *Machine learning: A probabilistic perspective*. MIT press, 2012.

[23] Vladimir Nekrasov, Hao Chen, Chunhua Shen, and Ian Reid. Architecture search of dynamic cells for semantic video segmentation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1970–1979, 2020.

[24] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6819–6828, 2018.

[25] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9226–9235, 2019.

[26] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

[27] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.

[28] M. Rebol and P. Knöbelreiter. Frame-to-frame consistent semantic segmentation. *arXiv preprint arXiv:2008.00948*, 2020.

[29] Fitsum Reda, Robert Pottorff, Jon Barker, and Bryan Catanzaro. flownet2-pytorch: Pytorch implementation of flownet 2.0: Evolution of optical flow estimation with deep networks. https://github.com/NVIDIA/flownet2-pytorch, 2017.

[30] R. Sibechi, O. Booij, N. Baka, and P. Bloem. Exploiting temporality for semi-supervised video segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[31] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[32] Yan Tian, Yujie Zhang, Di Zhou, Guohua Cheng, Wei-Gang Chen, and Ruili Wang. Triple attention network for video segmentation. *Neurocomputing*, 417:202–211, 2020.

[33] Serin Varghese, Yasin Bayzidi, Andreas Bar, Nikhil Kapoor, Sounak Lahiri, Jan David Schneider, Nico M Schmidt, Peter Schlicht, Fabian Huger, and Tim Fingscheidt. Unsupervised temporal consistency metric for video segmentation in highly-automated driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 336–337, 2020.

[34] 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*, 2020.

[35] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. Ranet: Ranking attention network for fast video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3978–3987, 2019.

[36] Rui Yao, Guosheng Lin, Shixiong Xia, Jiaqi Zhao, and Yong Zhou. Video object segmentation and tracking: A survey. *arXiv preprint arXiv:1904.09172*, 2019.

[37] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 2016.

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