# Leverage Temporal Consistency for Robust Semantic Video Segmentation

**Timo Sämann** [1] **Karl Amende** [1] **Stefan Milz** [1] **Horst Michael Gross** [2]

## Abstract

The majority of semantic segmentation models are unable to use temporal consistency in video data. This leads to the fact that short-term perturbations in the input data can cause erroneous predictions, which can have fatal consequences in safety-critical applications such as autonomous driving. We present an approach that integrates the temporal consistency of video data as a priori knowledge into the model. We achieve a much more robust semantic segmentation in case of perturbations in the input data. Furthermore, our approach improves the semantic segmentation for input data that does not contain perturbations. In both cases we demonstrate the qualitative and quantitative advantages of our approach.

## 1. Introduction

The realization of highly automated driving requires the intensive use of deep learning methods. One of the major challenges when using deep learning methods in the automotive industry are the high safety requirements for the algorithms. The use of black box solutions causes a potential safety risk and is therefore not permitted. For this reason, deep learning methods are needed that show comprehensible behaviour for us humans. In addition, the algorithms must be reliable and robust in the case of perturbations in the input data. These perturbations can be caused by sensor errors, external contamination of the sensors, overexposure or the occurrence of adversarial examples. Objects that suddenly appear or disappear from one frame to another due to inaccurate prediction or occurring perturbations can have disastrous consequences. These aspects receive less attention in the scientific community and are neglected in public data sets.

One way to achieve robustness against perturbations is to use temporal consistency in video data. The vast majority of previous deep neural networks have an independent single image prediction of the currently recorded scene, i.e. the same operations are performed for all input images and the information already received from the previous time step is discarded. For video processing in which there is a temporal continuity of the image content, the use of the information from previous time steps can overcome perturbations that would otherwise lead to miss-classification.

With our approach, we are able to overcome perturbations by incorporating the relevant information from earlier time steps into the current prediction. The idea is to combine the calculated feature maps from earlier time steps with the current feature map to compensate for shortcomings in the current prediction. This requires warping the feature maps from previous time steps $t_{-1}$ to $t_{-n}$ into the time stage $t_0$, where $n$ is the number of previous time steps. Warping takes place via the optical flow, following the idea of (Gadde et al., 2017). According to our experiments, a naive combination of the complete feature maps does not always lead to an improvement of the results. There are two main reasons for this:

1. It is in the nature of things that frames from previous time steps are less relevant than the current frame. Objects that appear in the image for the first time, e.g. because they have been covered by another object, cannot be represented by warping.

2. The warping process depends on the quality of the optical flow. Especially objects with a low pixel density like pole where the optical flow is not precise enough suffer in quality.

Therefore, a confidence-based combination of feature maps is performed that significantly reduces these issues. The confidence map gives us a confidence value for each pixel in the image that estimates the confidence of the prediction. The confidence map is obtained by probabilities from softmax distributions, which we have calibrated to obtain a reliable confidence estimate. We have observe that the confidence maps have a relatively low value at the areas in the image where we have inserted a perturbation, cf. (Hendrycks & Gimpel, 2016). Therefore, we use the confidence values as a measure of which areas of the feature maps $t_{-1}$ to $t_{-n}$ we

[1]Valeo Schalter und Sensoren GmbH, Kronach, Germany [2]Technical University Ilmenau, Ilmenau, Germany. Correspondence to: firstname.lastname <@valeo.com>, Horst Michael Gross <horst-michael.gross@tu-ilmenau.de>.

Presented at the ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning. 

combine with the feature map $t_0$. For the combination, a weighting is used that can be derived from the confidence values of the current and previous confidence maps. The areas of feature maps that have a higher confidence than the areas of the current feature map are combined. The combined feature map $fm\_new_{t0}$ then serves as the new feature map $fm_{t_{-1}}$.

To demonstrate the effectiveness of our approach, we use semantic video segmentation applied to two test data sets: One set of test data with artificially added perturbations, such as image artifacts, masking and adversarial pattern. And another one with the same images, but without any perturbations. We show that our approach not only significantly outperforms the perturbed data set but also slightly improves the baseline on the *clean* data set. Our approach is independent of the network architecture and does not require any further training or fine-tuning.

## 2. Pipeline

The entire pipeline of our approach is shown for a single time step $t_0$ in Fig. 1. As an example DNN we use the ENet architecture (Paszke et al., 2016), a ResNet based network that has a very low runtime while providing a respectable quality. However, the model can be exchanged with any other architecture. The model was trained on an internal fisheye training data set with 17 classes. After the last layer, the number of feature maps is 17 (referenced as $fm_{t0}$ in Fig. 1). The calculation of the argmax[1] and the following coloring lead to the baseline of semantic segmentation for the current time step, which is in Fig. 1 referred as $Segmentation_{t0}$.

The confidence map of the current time step $cm_{t0}$ is determined by the probabilities from softmax distributions. To obtain reliable confidence values, the confidence values are calibrated by modifying the softmax layer (see subsection 3). Furthermore, we compared these confidence maps with the epistemic uncertainty obtained by Monte Carlo Dropout (Kendall & Gal, 2017). It has turned out that the difference is usually fairly small, so we consider the softmax as an runtime efficient alternative.

The confidence and feature maps are warped in a so-called *warp* module (see box with red border in Fig. 1). The function of the *warp* module is to warp the feature or confidence maps from past time steps into the current time step in order to obtain a aligned representation. For warping we use the optical flow that we create with FlowNet2 (Ilg et al., 2017). Please note that we apply this model to fisheye camera images, although the model was trained on pinhole camera images. This leads to slightly worse results at the lateral areas of the image.

[1]For every pixel, the index of the maximum value along the depth axis is determined.

This aligned confidence maps are processed in the so-called *thresh* module (see box with green border in Fig 1) with threshold values and a weighting. The resulting confidence maps can be considered as a mask that is used for multiplication with the feature maps in the *combine* module (see box with blue border in Fig 1). In the *combine* module, the feature maps from the warp module are multiplied by the threshold confidence maps from the *thresh* module. The output of the combine module are 17 feature maps, which are composed pixel by pixel from the feature maps of time steps $t_0$ to $t_{-n}$. The new confidence map is called $cm\_new_{t0}$ and the robust semantic segmentation $RobustSegmentation_{t0}$. Please note that all results in section 5 refer to $n = 2$.

## 3. Confidence Calibration

Confidence calibration describes the problem of predicting probability estimates that are representative of the true probability of correctness (Guo et al., 2017). In other words, the aim of confidence calibration is to achieve the best possible consistency in predicting confidence and accuracy. For example, if the confidence of an image results in 90%, the accuracy of this image should also result in 90%. (Guo et al., 2017) has found that modern networks tend to be overconfidence in predicting confidences. The reason for the overconfidence of modern networks is the increased network capacity, the use of batch normalization and weight decay. A metric that indicates how well the network is calibrated is the Expected Calibration Error (ECE). To the best of our knowledge, the ECE metric has so far only been applied for image classification. In contrast to image classification, in semantic segmentation we do not calculate the gap between $acc$ and $conf$ per image but per pixel. This change requires an additional loop over all images that average the ECE. More formally, we describe the ECE for semantic segmentation as

$$\mathrm{ECE} = \sum_{l=1}^{L} \sum_{m=1}^{M} \frac{\|B_{m;l}\|}{n} \left\|\mathrm{acc}(B_{m;l}) - \mathrm{conf}(B_{m;l})\right\| \quad (1)$$

with

$$\mathrm{acc}(B_{m;l}) = \frac{1}{\|B_{m;l}\|} \sum_{i \in B_{m;l}} \mathbf{1}(\hat{y}_i = y_i) \quad (2)$$

$$\mathrm{conf}(B_{m;l}) = \frac{1}{\|B_{m;l}\|} \sum_{i \in B_{m;l}} \hat{p}_i. \quad (3)$$

We designate $L$ as the number of images, $M$ as the number of interval bins in which the predictions are grouped[2], $B$ as the number indices of pixels whose predictions (accuracy

[2]The number of interval bins in our case is 5, which leads to a bin size of 20% (100%/5) each.

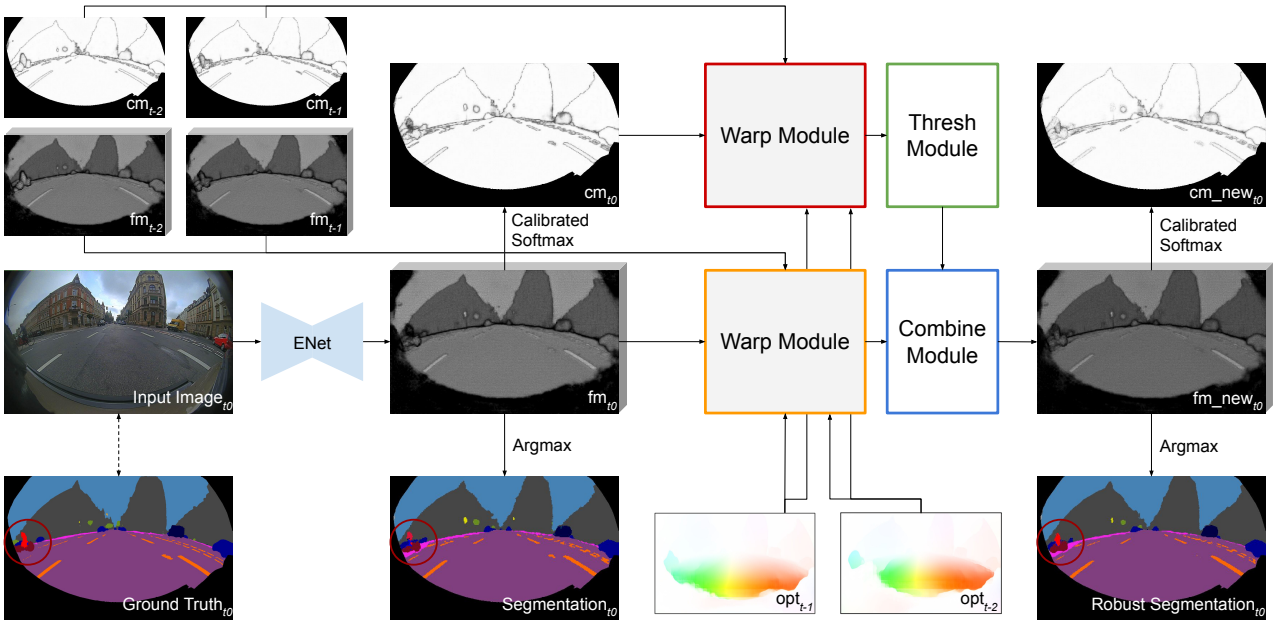


*Figure 1.* Overview of our pipeline. The abbreviations "cm", "fm" and "opt" stand for confidence map, feature map and optical flow respectively. White areas in the cm mean a high and black a low confidence.

and confidence respectively) falls into the bin interval, $n$ as the number of pixel per image, $\hat{y}$ as the prediction of the class, $\hat{p}$ as the prediction of the confidence, $y$ as ground truth and $i$ as the number of pixel per image and bin.

To calibrate the DNN we use temperature scaling, which consists of a single value added to the softmax layer. It was found in (Guo et al., 2017) that this type of calibration is the simplest and most effective at the same time. The extension provides for a division of the input of the softmax layer $z$ with a scalar $T$ (see Eq. 4). The optimal temperature scaling parameter was determined by Grid Search on our validation data set. This allowed us to reduce the ECE from 1.9 to 1.1. Relevant reference values from the literature could not be found. A direct comparison with ECE values from the task of image classification is not possible since the calculation is different.

$$\text{softmax} = \frac{e^{\frac{z}{T}}}{\sum_i e^{\frac{z_i}{T}}} \tag{4}$$

## 4. Data set

In order to test our approach, we have created our own data set consisting of a sequence of 1200 images which we refer to as *clean* data set. The images were taken with a fisheye camera and show scenes from the downtown area of a German city. To generate ground truth data we use the Tu-Simple-DUC model (Wang et al., 2018), which we trained on a fisheye camera data set before. In order to test the robustness of our network, a second data set was created by adding perturbations to the first data set, which we refer to as *perturb* data set. The perturbations can be divided into 3 different categories: Random patterns (random changes of multiple color channels), real perturbations (e.g. caused by packet loss) and adversarial patterns (generated from (Xie et al., 2017)). Images to which a perturbation was assigned were selected at random. Furthermore, the perturbations were placed at random locations in the image and can occur up to 6 times per image. In addition, perturbations also occur over several frames to evaluate robustness over a longer lasting perturbations. In total, 412 (33.67%) of 1200 images contain at least one added perturbation.

## 5. Results

We evaluate our approach qualitatively and quantitatively on the basis of two data sets: One without added perturbation pattern, which we call *clean*, and one with which we call *pertub*, see section 4. For qualitative evaluation we use the mean intersection over union (mIoU) and the global accuracy. The mIoU for the *clean* data set could be clearly improved from 62.39% to 63.20%. The IoU values per class are listed in Table 1. Apart from the classes *pole, traffic light* and *rider*, the values have increased significantly. One reason for the deterioration of these classes can be found in the inaccurate optical flow. A correct warping of the class *pole* requires a very precise optical flow. The global

accuracy, which indicates the percentage of pixels correctly classified, could be increased from 95.31% to 95.56%. Evaluated on our *perturb* data set the baseline worsens to a mIoU of 57.51% and a global accuracy of 93.87%. With our approach we achieve a significant increase of the mIoU from over 2.3% to 59.86% and a global accuracy of 94.61%. Due to the low confidence values at the locations of the perturbation patterns, these locations are used for combination. In this way, the negative effects of perturbations on prediction can be overcome or mitigated.

The supplementary material contains Fig. 2, 3, 4 and 5 that show the qualitative results for the data set *clean* and *perturb*. Four images are viewed in consecutive time steps. (a) represents the input image, (b) the pseudo ground truth (see section 4), (c) the baseline and (d) our approach. With our approach we achieve a much more stable and robust prediction in Fig. 2 and 3. Please note that our approach generally looks much smoother than the baseline, although the resolution is exactly the same. Even more clearly, the improvement can be seen in Fig. 4 and 5 for our *perturb* data set. Please note the image caption for further information.

*Table 1.* Quantitative results of our *clean* (abbr. "cle") and *perturb* data set (abbr. "per"). Comparison of the baseline (abbr. "Base") and our approach (abbr. "Ours"). All values are given in percent and indicate the IoU.

| Classes | Base-cle | Ours-cle | Base-per | Ours-per |
|---|---|---|---|---|
| Road | 95.66 | **95.90** | 94.59 | **95.39** |
| Sidewalk | 73.13 | **74.06** | 70.19 | **72.42** |
| Building | 92.36 | **92.71** | 89.97 | **90.83** |
| Wall | 64.96 | **68.03** | 43.57 | **52.02** |
| Fence | 20.26 | **20.99** | 17.72 | **18.57** |
| Pole | **39.11** | 38.16 | **37.50** | 36.74 |
| Traffic light | **47.86** | 47.26 | **46.32** | 45.50 |
| Traffic sign | 48.32 | **49.98** | 46.15 | **48.08** |
| Vegetation | 85.30 | **85.86** | 79.47 | **81.26** |
| Terrain | 31.90 | **33.31** | 23.87 | **26.58** |
| Sky | 96.10 | **96.35** | 94.46 | **95.18** |
| Person | 43.98 | **45.40** | 42.39 | **44.21** |
| Rider | **40.13** | 39.12 | 29.81 | **37.22** |
| Car | 86.75 | **87.07** | 82.46 | **84.17** |
| Truck | 81.84 | **82.87** | 73.92 | **78.31** |
| Bicycle | 46.16 | **48.66** | 45.03 | **47.65** |
| Road markings | 66.89 | **68.73** | 60.17 | **63.49** |
| **Mean IoU** | 62.39 | **63.20** | 57.51 | **59.86** |

## 6. Conclusion and Future Work

Safety-critical applications require reliable and robust algorithms. We introduced an approach that allows a DNN for semantic image segmentation to leverage consistency in video data to make the prediction much more robust. With regard to suddenly occurring perturbations in the input data, our approach can drastically increase the robustness of the prediction. But even under normal conditions a more stable prediction can be achieved, which we have shown qualitatively and quantitatively. We see considerable potential for improvement in our approach through better uncertainty modeling. The knowledge of the exact localization of the image regions where the DNN is uncertain is a crucial point for the effectiveness of our approach. For this reason we plan to replace the calibrated probabilities from softmax distributions with different types of uncertainty modelling.

## References


Gadde, R., Jampani, V., and Gehler, P. V. Semantic video cnns through representation warping. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4453–4462, 2017.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2462–2470, 2017.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pp. 5574–5584, 2017.

Paszke, A., Chaurasia, A., Kim, S., and Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., and Cottrell, G. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1451–1460. IEEE, 2018.

Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., and Yuille, A. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1369–1378, 2017.


## A. Supplementary Material

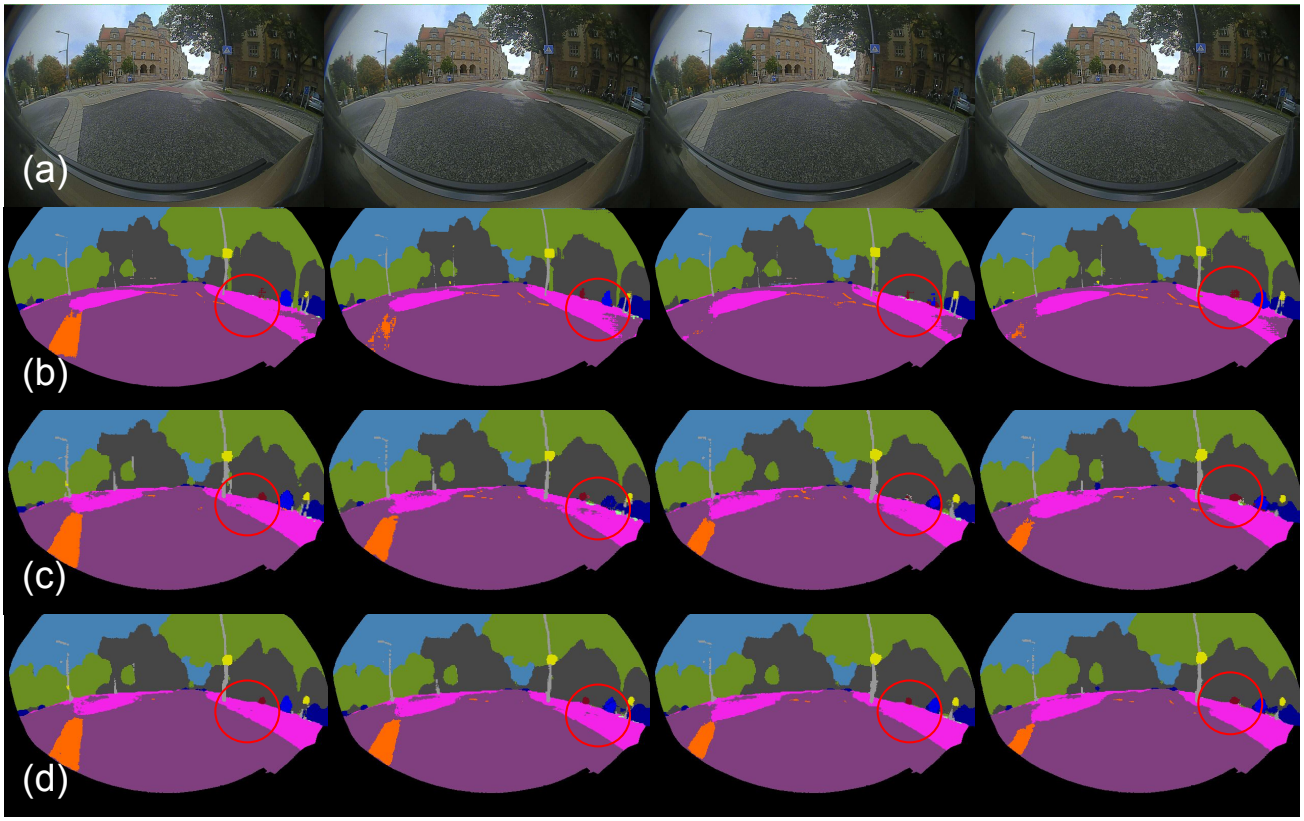


*Figure 2.* Qualitative results from our *clean* data set. 4 images are shown in consecutive time steps. (a) input image, (b) pseudo ground truth, (c) baseline, (d) our approach. It can be seen that the sidewalk in all pictures is much denser and has fewer holes. Furthermore, the baseline shows a class change between *motorcycle* and *car* in column 2, as well as the disappearance of *bicycle* in columns 3, which does not happen with our approach.

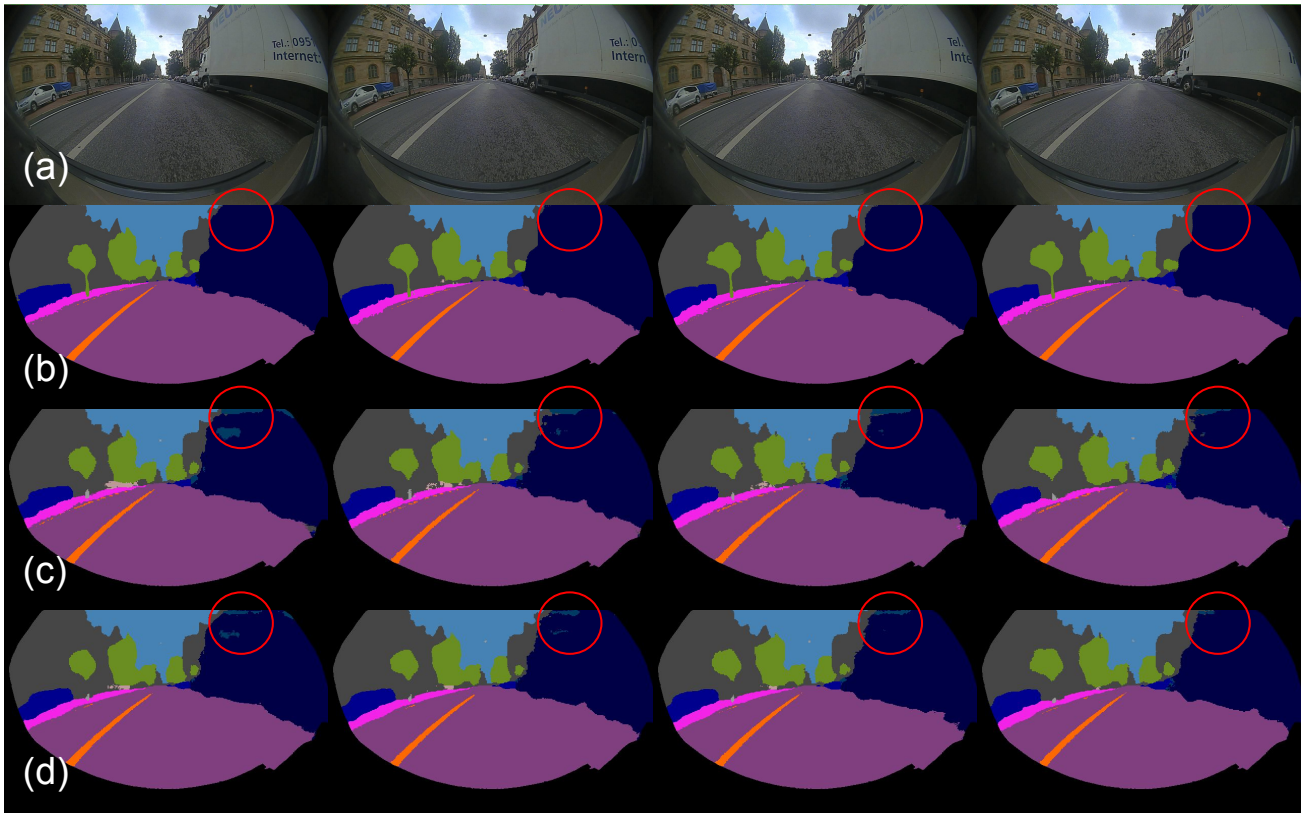


*Figure 3.* Qualitative results from our *clean* data set. 4 images are shown in consecutive time steps. (a) input image, (b) pseudo ground truth, (c) baseline, (d) our approach. The class *truck* is predicted much more stable compared to the baseline.

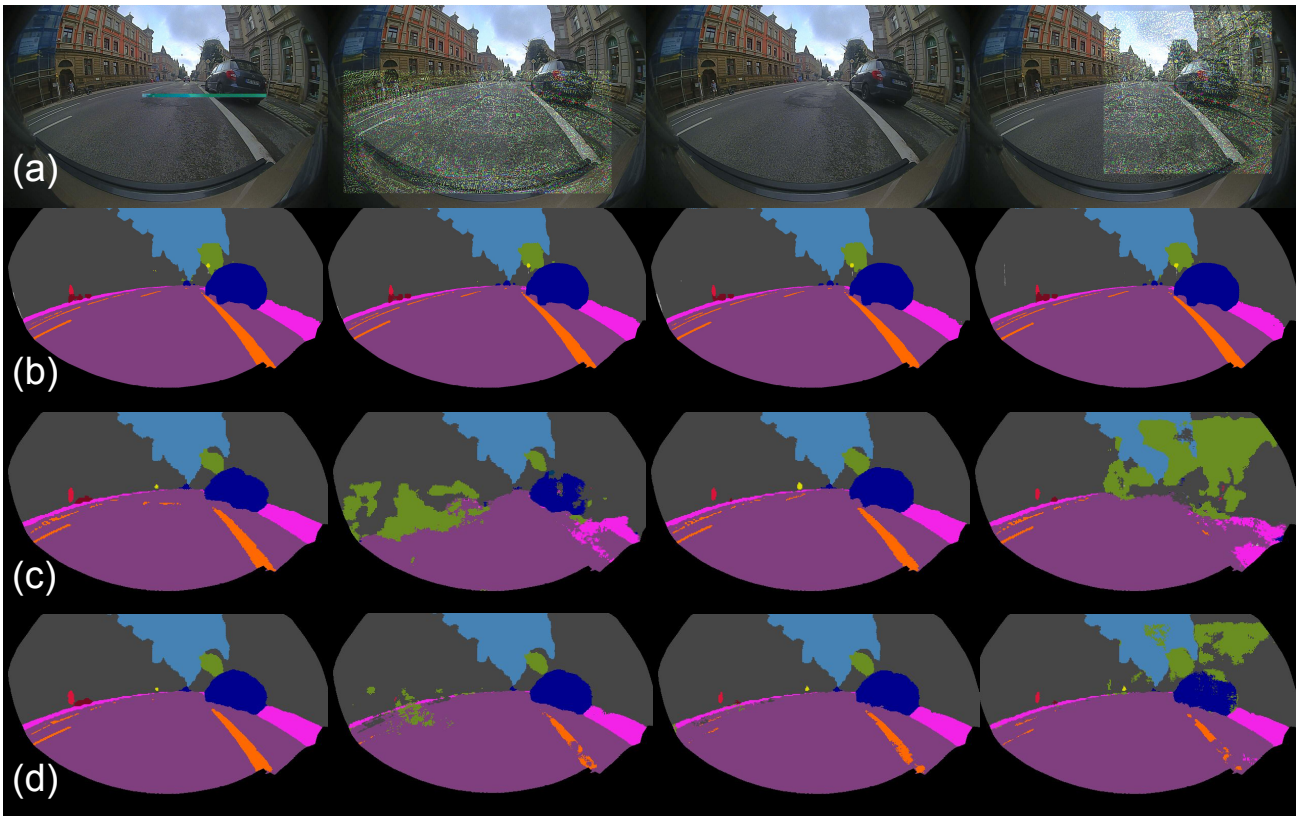


*Figure 4.* Qualitative results from our *perturb* data set. 4 images are shown in consecutive time steps. (a) input image, (b) pseudo ground truth, (c) baseline, (d) our approach. The perturbation pattern in columns 2 and 4 drastically destroys the prediction of the baseline, while our approach drastically reduces the influence of the perturbation on the prediction. Please note that the perturbation patterns in the image are amplified for visualization reasons.

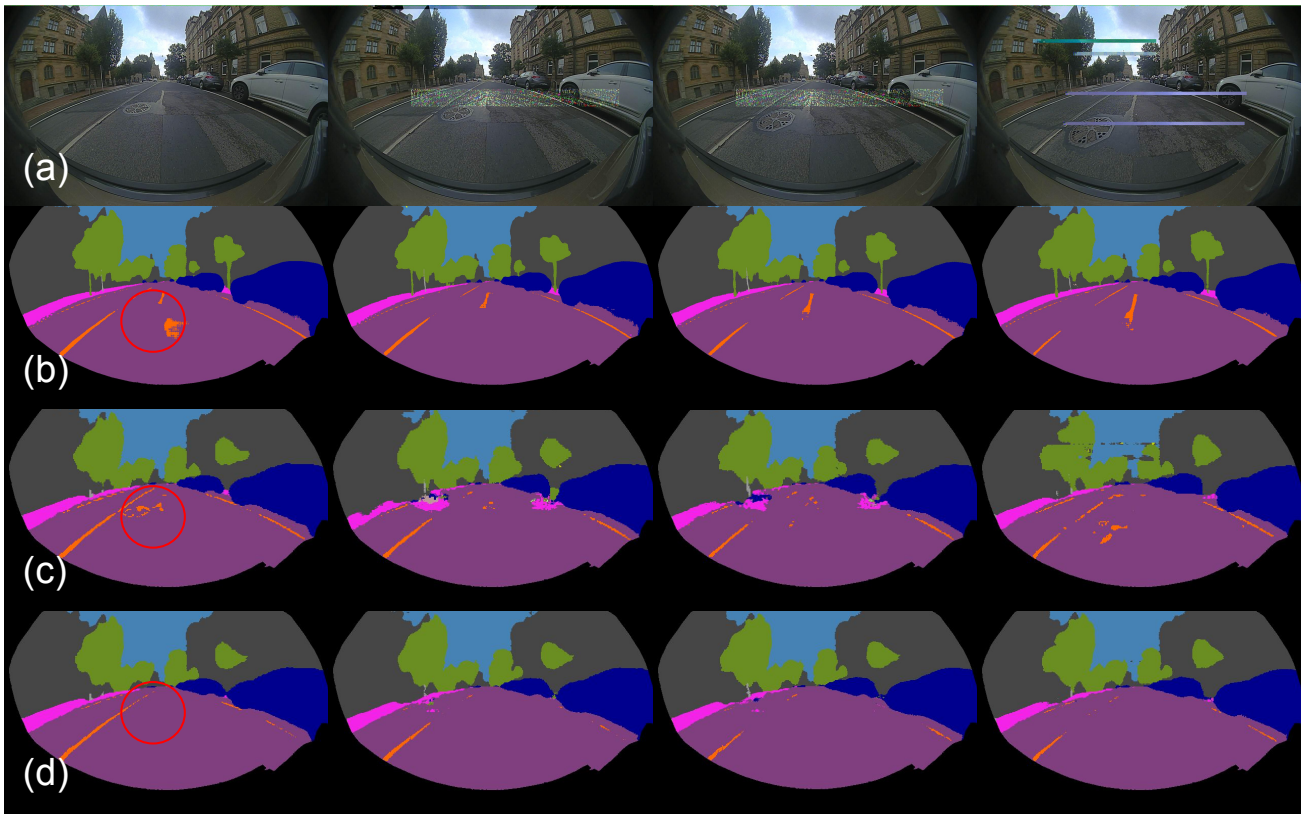


*Figure 5.* Qualitative results from our *perturb* data set. 4 images are shown in consecutive time steps. (a) input image, (b) pseudo ground truth, (c) baseline, (d) our approach. In the first column the class *road marking* is wrongly detected by the baseline and the ground truth. In the other columns it can be seen that the perturbations in the input data affect our approach much less. Please note that the perturbation patterns in the image are amplified for visualization reasons.