Foveal Vision for Instance Segmentation of Road Images

Benedikt Ortelt[1], Christian Herrmann[2,3], Dieter Willersinn[2] and Jürgen Beyerer[2,3]

[1] *Robert Bosch GmbH, Leonberg, Germany*

[2] *Fraunhofer IOSB, Karlsruhe, Germany*

[3] *Karlsruhe Institute of Technology KIT, Vision and Fusion Lab, Karlsruhe, Germany*

Keywords: Instance Segmentation, Multi-scale Analysis, Foveated Imaging, Cityscapes.

Abstract: Instance segmentation is an important task for the interpretation of images in the area of autonomous or assisted driving applications. Not only indicating the semantic class for each pixel of an image, but also separating different instances of the same class, even if neighboring in the image, it can replace a multi-class object detector. In addition, it offers a better localization of objects in the image by replacing the object detector bounding box with a fine-grained object shape. The recently presented Cityscapes dataset promoted this topic by offering a large set of data labeled at pixel level. Building on the previous work of (Uhrig et al., 2016), this work proposes two improvements compared to this baseline strategy leading to significant performance improvements. First, a better distance measure for angular differences, which is unaffected by the $-\pi/\pi$ discontinuity, is proposed. This leads to improved object center localization. Second, the imagery from vehicle perspective includes a fixed vanishing point. A foveal concept counteracts the fact that objects get smaller in the image towards this point. This strategy especially improves the results for small objects in large distances from the vehicle.

1 INTRODUCTION

Understanding the scene in road images is important for assisted and autonomous driving applications. Information about driving related aspects, such as road boundaries, object types, free space or obstacles, is of particular interest. Given images from vehicle perspective, an understanding of the surrounding environment is required. Instance segmentation serves this purpose by denoting for each image pixel the underlying object instance and the object class (e.g., car, truck or pedestrian). Compared with regular semantic segmentation, neighboring instances of the same class, e.g., two vehicles, are separated, which is important in traffic situations where each object might behave differently. In this matter, the instance segmentation replaces an object detector and additionally provides a pixel-level mask for each object instead of a simple bounding box. This allows a better localization of the object.

Following the widespread terminology (He et al., 2017), *semantic segmentation* denotes the per-pixel classification of the image content, *object detection* the acquiring of object bounding boxes and *instance segmentation* the per-pixel indication of class and instance which makes this some kind of combination of segmentation and detection. While semantic segmentation is a very popular research field, instance segmentation receives less attention. This can also be observed by the number of submitted results for the popular Cityscapes benchmark[1] where regular semantic segmentation leads with 52:13 submissions at the time of writing. Nevertheless, the Cityscapes dataset (Cordts et al., 2016) boosted the instance segmentation topic a lot by providing appropriately annotated data.

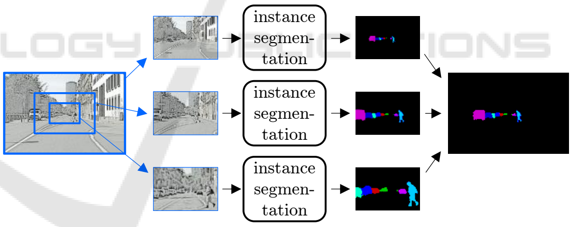

Figure 1: Foveal image analysis concept.

This work builds upon the prior work of Uhrig et al. (Uhrig et al., 2016) and improves their solution by two specific contributions. First, an improvement

[1] https://www.cityscapes-dataset.com/benchmarks

Ortelt, B., Herrmann, C., Willersinn, D. and Beyerer, J.
Foveal Vision for Instance Segmentation of Road Images.
DOI: 10.5220/0006616103710378
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 4: VISAPP*, pages 371-378
ISBN: 978-989-758-290-5

which is specific to this approach by replacing the distance measure for angular differences with an improved version that is invariant to the $-\pi/\pi$ discontinuity. Second, a generic multi-scale foveal concept is proposed to compensate for the perspective properties of road image data where objects tend to become smaller in the image with increasing distance. A multi-scale fine-grained analysis around the vanishing point of the road allows to better address these small objects as illustrated in figure 1.

2 RELATED WORK

The success of object detection approaches (Girshick, 2015; Ren et al., 2015; Liu et al., 2016) gave rise to the fine-grained detection of object contours in the shape of instance segmentation. Two principle concepts to address instance segmentation can be distinguished. First, instance candidate approaches build upon the reliable bounding boxes of an object detector to refine the object shape. Second, candidate free methods circumvent the object detection by direct estimation of the object's shape. Regarding dataset choices, while Cityscapes (Cordts et al., 2016) is the current state-of-the-art benchmark for instance segmentation in road images, the MS COCO dataset (Lin et al., 2014) offers a wider spectrum of image contents and the KITTI dataset (Geiger et al., 2012) with additional instance labels (Chen et al., 2014; Zhang et al., 2016a) is an older and smaller road image choice. Because this paper focuses on road scene understanding, the Cityscapes dataset will be used for training and testing the proposed methods.

2.1 Instance Candidate Methods

The refinement of bounding boxes allows to distinguish between instance pixels and background pixels. The most direct approach is to learn a mask-regression in addition to the bounding box regression and the class label (He et al., 2017). Fusing several sub-networks by a cascade (Dai et al., 2016) or using an identification network with a Conditional Random Field (CRF) (Arnab and Torr, 2017) are further options. Fully Convolutional Neural Networks (FCNs) (Long et al., 2015) can be used to improve erroneous bounding boxes by adjusting their size (Hayder et al., 2016).

2.2 Candidate Free Methods

Caused by the success of semantic labeling, methods solving instance segmentation on pixel-level are the second option. (Zhang et al., 2015; Zhang et al., 2016b) apply a FCN and exploit depth information for a stable training process to generate instance candidates which are refined by a Markov Random Field (MRF) which reduces the error by enforcing global consistency. (Romera-Paredes and Torr, 2016) propose a recurrent solution where instances of one class are separated step by step. Another solution for instance segmentation is to start from a semantic segmentation result where predicting the contour of an object allows to separate single instances of one class (Ronneberger et al., 2015; van den Brand et al., 2016; Kirillov et al., 2016). Such methods are difficult to train and have issues with split instances, e.g., because of occlusion. This is addressed by (Uhrig et al., 2016) by additionally predicting the object distance and the direction to the respective object center for each pixel. A post-processing step allows then to detect and merge split instances. This paper builds upon this approach which will be elaborated on in more detail in the next section.

2.3 Multi-scale and Foveated Concepts

Multi-scale image processing, e.g., by image pyramids, is a widespread technique to address differently sized objects in images. It is a key part of all popular object detection frameworks (Viola and Jones, 2004; Ren et al., 2015; Liu et al., 2016). While necessary to achieve good results, it can introduce a significant computational overhead. Foveated imaging is a concept which reduces this additional burden by spatially restricting the higher resolution scales around a fixation point (Bandera and Scott, 1989; Ude et al., 2003; Wang and Bovik, 2006). An increasing resolution is gradually compensated by limitation to a smaller spatial region, which imitates the human vision.

3 BASIC CONCEPT

The proposed method builds upon (Uhrig et al., 2016). This approach performs instance segmentation in two stages. In the beginning, an extended FCN8s (Long et al., 2015), based on VGG16 (Simonyan and Zisserman, 2015), is used for CNN-based pixelwise predictions of the semantic class, the depth and the instance-based geometric direction to the instance center, as shown at the top of figure 2.

These three prediction maps are then post-processed to obtain the final instance segmentation. The semantic label is utilized to distinguish between instances of different semantic classes and to classify the resulting instances in the end while the depth la-

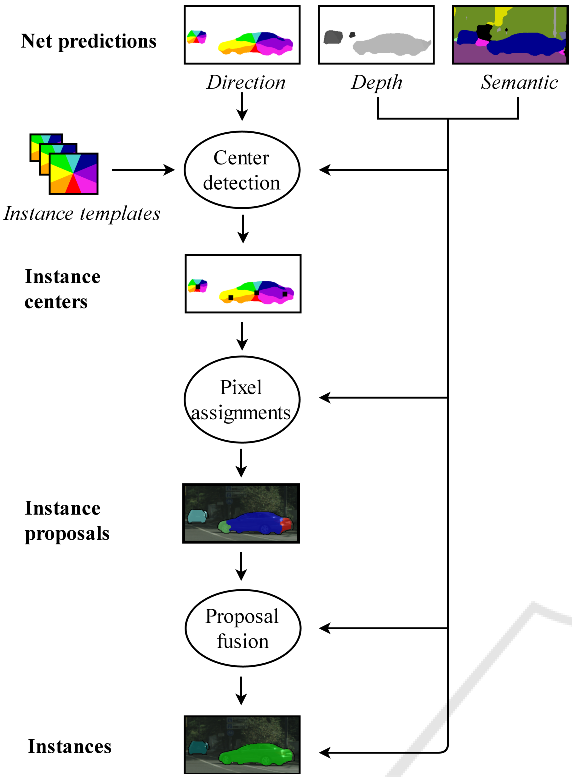

Figure 2: The post processing concept of (Uhrig et al., 2016).

bel is mainly used to estimate each instance's shape and size. The direction predictions offer the direction to the corresponding instance center of every instance pixel which yields a high discriminativity at instance borders.

In the post-processing, the predicted maps are jointly employed to generate the instances by finding the instance centers and assigning all pixels to the appropriate centers. To reduce oversegmentation caused by wrong extra instance centers, instances are fused subsequently depending on their properties. Initial instance centers are determined with a template matching procedure on the direction map using the Normalized Cross Correlation (NCC)

$$R(x,y) = \frac{\sum_{x',y'} \mathbf{F}(x+x',y+y') \cdot \mathbf{T}(x',y')}{\sqrt{\sum_{x',y'} \mathbf{F}(x+x',y+y')^2 \cdot \sum_{x',y'} \mathbf{T}(x',y')^2}} \qquad (1)$$

to compare the angular patterns between the direction map $\mathbf{F}$ and the template $\mathbf{T}$. The rectangular templates are scaled according to prior knowledge and the information from the semantic and depth map. Maxima in the score maps correspond to instance centers that are found with non-maximum suppression in a region that is equivalent to the template area. Pixels are assigned to the closest center where semantic, predicted direction and relative location agree which results in instance proposals. Two proposals are fused afterwards if one proposal's accumulated direction is biased to a neighboring proposal with similar depth and semantic. We refer to (Uhrig et al., 2016) for further details.

4 IMPROVED ANGULAR DISTANCE

The comparison of the template and the directions map by NCC sometimes leads to inaccurate initial center detections. This includes spatial shifts, missed centers or even multiple centers per instance. These effects may lead to inaccurate instance localization, missed instances or over-segmented instances.

For this reason, we suggest a more robust and precise Improved Angular Distance (IAD) to compare the angular patterns for detecting the instance centers. In contrast to the basic NCC approach, it is invariant to the angle's non-linearity between $-\pi$ and π.

Instead of the scalar angles, comparison of the center directions is performed directly by the direction vectors in shape of vector fields. For a single pixel, the network predictions p_i for each discrete angular range class are used as weights for each class's respective direction vector $\boldsymbol{r}_i$ to get the final direction vector (Uhrig et al., 2016)

$$\rho = \sum_i p_i \cdot \boldsymbol{r}_i \,. \qquad (2)$$

Consequently, the template $\mathbf{T}$ and the direction field $\mathbf{F}$ are extended to vector fields containing the two dimensional normalized direction vectors $\hat{\rho}$ at each location.

For a template and a direction field region of height h and width w, the score S that indicates a pixel's likelihood of being an instance center is then

$$S(x,y) = \frac{1}{h \cdot w} \sum_{x',y'} \mathbf{F}(x+x',y+y')^\top \cdot \mathbf{T}(x',y'). \qquad (3)$$

This corresponds to the cross correlation for the three-dimensional case normalized by the number of pixels. The normalization allows the comparison of scores of different templates in the non-maximum suppression step.

Leaving the pixel normalization aside, the computation is equal to the inner product of the direction vectors $\boldsymbol{f}$ and $\boldsymbol{t}$ of each pixel. Because of vectors having unit length, this equals the cosine of the angle between the vectors

$$\langle \boldsymbol{f}, \boldsymbol{t} \rangle = \cos(\angle(\boldsymbol{f}, \boldsymbol{t})). \qquad (4)$$

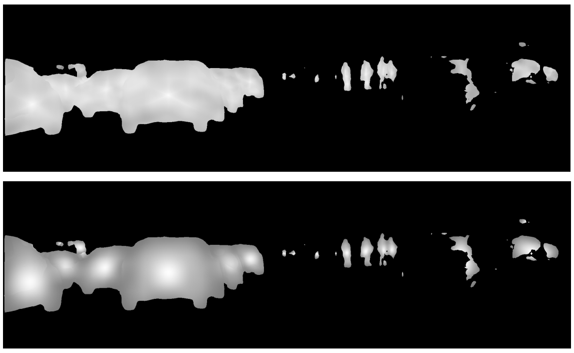

Figure 3: Score maps S when using NCC (top) and IAD (bottom).

For this reason, small angular differences have a small impact while opposed vectors cause significant score reductions. So, lateral shifts of centers are punished harder leading to more accurate results in comparison to NCC. Additionally, this leads to more distinct maxima in the score maps which are depicted in figure 3.

Due to normalization and the cosine function, the overall scores ranges from -1 to 1 where a score of 1 indicates a perfect match.

There are several options to further improve the score in equation 3. To better adjust the rectangular template to the instance shape, pixels classified as background can be omitted from score computation. Thereby, the rectangular template is adapted to the real shape of the instance. Effects at the borders between instances where opposing directions occur can be handled in two ways. The easiest way is to reduce the instance template size to avoid that neighboring instance pixels fall inside. We found that IAD is more robust to smaller templates than NCC. Addressing border effects on score level is possible by ignoring negative cosine values which indicate opposing directions occurring at the border.

5 FOVEAL STRATEGY

Foveated imaging is an intriguing but often unpopular concept because of two common downsides for general image analysis. First, it is often unclear how to choose the fixation point where resolution should be the highest. Either there is no clear point or knowing the fixation point already solves the targeted problem, which would be the case, e.g., in object detection. Second, applying foveal image transformations, such as a log-polar transform (Schwartz, 1977), on the input image alters object proportions depending on the object location in the image. This makes the training of a unified detection or segmentation method very difficult.

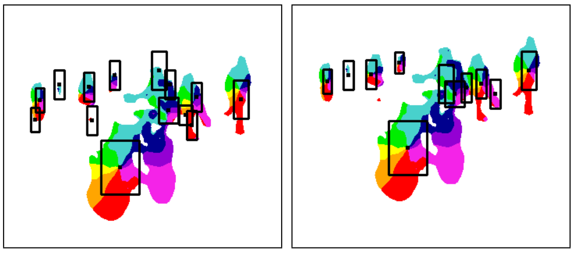

Figure 4: Comparison of instance center detections with NCC (left) and IAD (right). Color overlay for directions is based on (Uhrig et al., 2016).

Both issues are addressed in this work. First, the vanishing point of the road offers an obvious choice as fixation point in imagery captured in driving direction from a vehicle. Second, we explicitly extract image regions and scales instead of applying a highly nonlinear foveal image transformation to the input image. This can also be understood as a multi-scale image pyramid strategy with high resolution scales being spatially focused and limited to the fixation point.

Two different solutions to determine the fixation point in the road images are suggested:

1. Horizontally centered and vertically aligned at the pre-calibrated horizon at **fixed** image coordinates. This assumes a fixed and well aligned camera position. Extraction is quick.
2. **Dynamically** extracted based on the semantic segmentation output of the Convolutional Neural Network (CNN) at the original scale. This adapts the fixation point according to the currently analyzed scene but requires a small computational overhead.

The applied strategy to extract the dynamic fixation point is based on the road segmentation on the whole image. The topmost image area classified as road is considered the vanishing point of the road and consequently set as fixation point. This strategy results in an intersection over union of 0.693 between the cropped regions around the detected point and around the ground truth vanishing point of the road. Note that the CityScapes dataset also includes depth information from stereo vision which offers an additional way to determine the vanishing point. However, we decide for a strategy working also in the absence of depth information to broaden the scope of the approach.

5.1 Fusion Methods

Foveal regions are cropped in alignment with the fixation point as visualized in figure 5. By avoiding regions centered around the fixation point, instances are

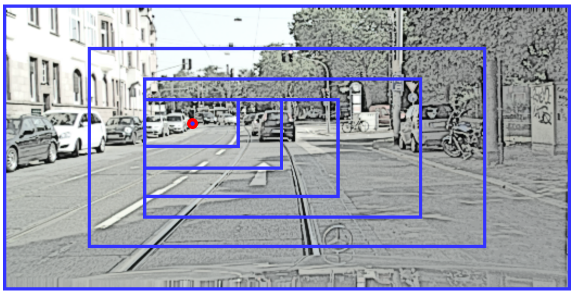

Figure 5: Foveal regions and their alignment with the fixation point.

better covered caused by the road beginning at the bottom center of the image.

For each crop, the region is rescaled to fit the input size of the instance segmentation CNN. Thus, small objects at larger distances are enlarged. This better exploits the typical scale range of objects from 40 to 140 pixels which networks pre-trained on the ImageNet data respond to (Hu and Ramanan, 2016). The CNN is fine-tuned separately for each foveal scale to adjust for the differing alignment. Experiments with a jointly trained network for all crops indicated worse results. The post-processing template sizes are adjusted accordingly to ensure consistency with the resized objects.

The proposed foveal strategy then results in an instance segmentation map for each crop, which has the original resolution. The segmentation results of the crops are subsequently scaled and merged into the result of the whole image beginning with the largest crop. Three different fusion methods to combine the crops are explored, with the first two being generally applicable to all instance segmentation approaches and the third one being specific to the chosen base approach.

Baseline Instance Fusion. As a baseline, a crop's instance segmentation map simply replaces the corresponding part in the larger crop's instance segmentation map. To avoid instances being cut at the crop border, instances that overlap within a small overlapping zone are merged.

Improved Instance Fusion. While smaller crops allow to better segment the smaller objects, it is likely that big instances are split because they significantly exceed the typical object size range that can be detected by the CNN. Additionally, there is less context that the CNN can rely on if only a part of the instance is visible in the crop. Following from this, the baseline instance fusion might insert these split instances even if they are correctly segmented in larger crops. This motivates an improved method to fuse the instance segmentation maps pursuing the objective of combining bigger instances segmented at smaller scales and smaller instances segmented at larger scales. Merging a smaller crop into a larger crop follows these steps:

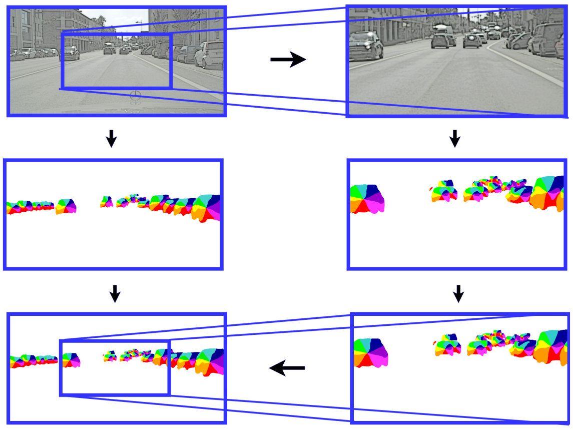

Figure 6: Illustration of the feature fusion concept.

1. Remove all instances in the larger crop that lie completely within the smaller crop region. Keep the remaining ones.
2. If more than 50% of an instance in the smaller crop overlap with a remaining instance from the larger crop
   (a) then add the instance's pixels to this remaining instance.
   (b) otherwise add the instance as a new one and overwrite any overlap with other instances.

This strategy keeps the larger instances from the larger crops while still exploiting the finer details in the smaller crops, both in terms of better object contours as well as small objects.

Feature Fusion. Both previous instance fusions are independent of the instance segmentation algorithm. Specific to the basic method of in this work, scale fusion can also be done earlier, i.e., on the semantic, depth and direction feature map. Because the post-processing is tolerant to minor inaccuracies, the different feature map crops are scaled accordingly and just copied into each other without any border effect handling. These fused maps are then regularly post-processed. Figure 6 illustrates this strategy.

6 EXPERIMENTS

The experiments are performed on the Cityscapes dataset (Cordts et al., 2016) following the official instance segmentation evaluation protocol. Instance segmentation results are thus given as average precision (AP) and AP$^{50\%}$, where AP$^{50\%}$ denotes the average precision for a fixed instance overlap thres-

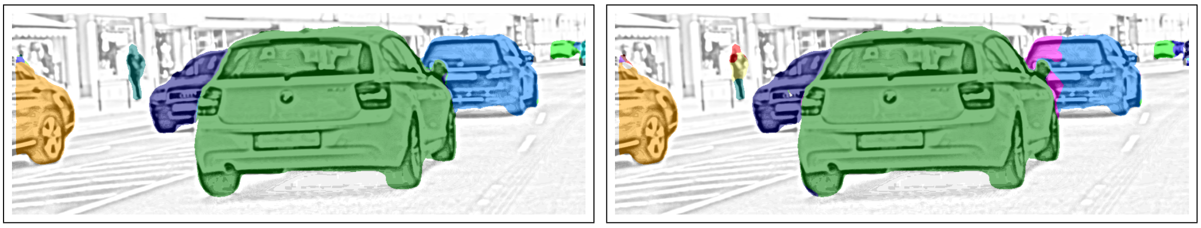

Figure 7: Comparison of instance segmentation improvement by IAD (left) compared with NCC (right).

Table 1: Evaluation of the IAD compared to the NCC baseline on Cityscapes *val*. Numbers in percent.

	NCC AP	IAD AP	NCC $AP^{50\%}$	IAD $AP^{50\%}$
person	7.7	**11.0**	20.9	**27.4**
rider	7.4	**8.8**	25.3	**26.9**
car	22.0	**24.5**	38.6	**40.6**
truck	8.3	**11.9**	17.1	**21.5**
bus	15.8	**17.3**	**35.5**	34.8
train	7.3	**10.1**	14.5	**22.0**
motorcycle	5.3	**7.0**	16.3	**19.0**
bicycle	5.8	**6.4**	19.0	**20.2**
mean	10.0	**12.1**	23.4	**26.6**

hold in terms of the Jaccard index (intersection over union) of 0.5. The CNN is trained on the training set containing 2,975 images. Method optimization is performed on the 500 image validation set. The test set with 1,525 images serves for the final evaluation where tests have to be performed via the official dataset website because no annotations are released for this set. Method validation will be performed against the reimplemented baseline of (Uhrig et al., 2016). The base network is trained the same way followed by the same post-processing strategy, which results in a comparable overall performance with an AP of 9.9 (theirs) versus 10.0 (ours) on the validation set.

6.1 Angular Distance

Replacing the NCC by the IAD in the post-processing for instance center detection, promises to improve the results because of the more robust and tolerant strategy. The results in table 1 clearly confirm this across all classes in the regular setting. Overall, results are improved from 10.0 to 12.1 percent. An example comparison of instance segmentation results between both options is depicted in figure 7. It indicates less decomposed instances for the IAD.

6.2 Multi-scale Analysis

Even though the Cityscapes dataset has a considerable image resolution of 2048 × 1024 pixels, far objects are mostly below 30 × 30 pixels in size. This motivates the proposed foveal strategy which enlarges the objects into the typical favorable range of the analyzing CNN. Due to the typical high response of ImageNet pre-trained networks to objects in the scale range of 40 to 140 pixels, an overly dense scale sampling of the cropped foveal regions is unnecessary. We decide for a scaling factor of 2 between crops, i.e., each region of interest has half the size of the previous one. The cropped regions are then upscaled to push object sizes into the favorable range of the CNN. Table 2 shows the results of the comparison between pre-calibrated and dynamically selected fixation points as well as an analysis of the number of useful crops. It shows that the dynamic selection becomes important if crops get smaller and more focused on small image regions. These regions must then be selected well to improve performance. When only using a single crop, the pre-calibrated strategy is sufficient because the vanishing point of the road is usually already inside the crop which renders the dynamic strategy unnecessary. Regarding the fusion strategy, feature fusion and the improved instance fusion are consistently superior to the baseline strategy. Also, the improved instance fusion outperforms the method specific feature fusion. Overall, the validation AP is increased from 12.1 to 14.4 percent.

Table 2: Comparison of fixation point search strategies, number of foveal crops and fusion strategy on Cityscapes *val*. All numbers are in percent and larger is better.

foveal setting	metric	feature fusion	baseline instance fusion	improved instance fusion
none	AP	12.1	-	-
fixed, 1 crop	AP	13.3	13.1	14.4
fixed, 2 crops	AP	8.9	8.4	13.8
dynamic, 1 crop	AP	12.6	12.5	14.1
dynamic, 2 crops	AP	11.1	10.9	14.0
none	$AP^{50\%}$	26.6	-	-
fixed, 1 crop	$AP^{50\%}$	29.3	28.2	29.8
fixed, 2 crops	$AP^{50\%}$	20.8	19.0	29.1
dynamic, 1 crop	$AP^{50\%}$	28.3	27.7	28.6
dynamic, 2 crops	$AP^{50\%}$	26.7	25.6	28.7

Having a look at instance segmentation results for different object categories in table 3 indicates that the improved instance fusion better addresses the large

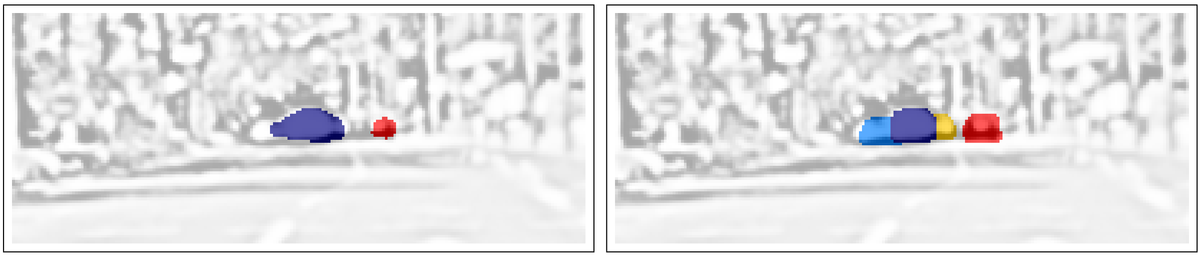

Figure 8: Result of the foveal strategy (left) on an image region where non-foveal processing detected no instances. The ground truth is given on the right for reference.

Table 3: Detailed results for selected foveal strategies on Cityscapes *val*. All numbers are in percent and larger is better.

foveal setting	fusion method	metric	person	rider	car	truck	bus	train	motorcycle	bicycle	mean
none	-	AP	11.0	8.8	24.5	11.9	17.3	10.1	7.0	6.4	12.1
fixed, 1 crop	improved instance fusion	AP	13.2	11.1	25.8	14.5	22.9	11.2	8.2	8.1	14.4
fixed, 1 crop	feature fusion	AP	14.4	11.7	27.8	8.7	19.0	9.1	8.2	7.7	13.3
dynamic, 1 crop	improved instance fusion	AP	12.9	10.7	25.1	14.0	23.7	10.3	8.1	7.7	14.1
dynamic, 1 crop	feature fusion	AP	13.9	11.7	26.5	7.8	17.9	7.6	8.4	7.1	12.6

Table 4: Instance-level segmentation results on Cityscapes *test*. Comparison of the published class-based performance of the basic concept (Uhrig et al., 2016) and our best performing method including IAD, the fixed foveal strategy with 1 crop and improved instance fusion. All numbers are in percent and larger is better.

	metric	person	rider	car	truck	bus	train	motorcycle	bicycle	mean
basic concept (Uhrig et al., 2016)	AP	12.5	**11.7**	22.5	3.3	5.9	3.2	6.9	5.1	8.9
Ours	AP	**13.4**	11.4	**24.5**	**9.4**	**14.5**	**12.2**	**8.0**	**6.7**	**12.5**
basic concept (Uhrig et al., 2016)	AP$^{50\%}$	**31.8**	**33.8**	37.8	7.6	12.0	8.5	**20.5**	17.2	21.1
Ours	AP$^{50\%}$	31.5	29.7	**40.0**	**16.0**	**23.8**	**21.7**	19.2	**19.9**	**25.2**
basic concept (Uhrig et al., 2016)	AP^{100m}	24.4	**20.3**	36.4	5.5	10.6	5.2	10.5	9.2	15.3
Ours	AP^{100m}	**24.5**	19.6	**39.3**	**14.5**	**24.2**	**18.5**	**11.1**	**11.1**	**20.4**
basic concept (Uhrig et al., 2016)	AP^{50m}	**25.0**	**21.0**	40.7	6.7	13.5	6.4	11.2	9.3	16.7
Ours	AP^{50m}	24.7	20.2	**42.5**	**17.2**	**27.6**	**21.8**	**11.7**	**11.3**	**22.1**

object classes as intended. By avoiding to split objects, such as trucks or trains, at crop borders, segmentation significantly improves. This is opposing to the feature fusion strategy where performance for these classes drops heavily caused by split instances.

Finally, the best-performing combination is compared with the published results of the basic method on the Cityscapes *test* dataset. Significant overall improvement is achieved. Only the results for the person and rider class show mixed results, indicating that the CNN used by (Uhrig et al., 2016) is significantly better at detecting people at the cost of all other classes.

7 CONCLUSION

Using an angular distance which is unaffected by the $-\pi/\pi$ discontinuity improved instance segmentation results significantly. Further performance progress was made by a generally applicable foveal image analysis strategy with a multi-scale focus on the vanishing point of the road. This allowed to better distinguish far objects which are typically small in the image. Overall, the performance on the Cityscapes test dataset was improved by both measures from 8.9 to 12.5 percent average precision.

REFERENCES

Arnab, A. and Torr, P. (2017). Pixelwise Instance Segmentation with a Dynamically Instantiated Network. In *Conference on Computer Vision and Pattern Recognition*.

Bandera, C. and Scott, P. D. (1989). Foveal machine vision systems. In *Systems, Man and Cybernetics, 1989. Conference Proceedings., IEEE International Conference on*, pages 596–599. IEEE.

Chen, L.-C., Fidler, S., Yuille, A. L., and Urtasun, R. (2014). Beat the mturkers: Automatic image labeling from weak 3d supervision. In *Conference on Computer Vision and Pattern Recognition*, pages 3198–3205.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Conference on Computer Vision and Pattern Recognition*.

Dai, J., He, K., and Sun, J. (2016). Instance-aware Semantic Segmentation via Multi-task Network Cascades. In *Conference on Computer Vision and Pattern Recognition*.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition*.

Girshick, R. (2015). Fast R-CNN. In *International Conference on Computer Vision*.

Hayder, Z., He, X., and Salzmann, M. (2016). Shape-aware Instance Segmentation. *arXiv preprint arXiv:1612.03129*.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. *arXiv preprint arXiv:1703.06870*.

Hu, P. and Ramanan, D. (2016). Finding tiny faces. *arXiv preprint arXiv:1612.04402*.

Kirillov, A., Levinkov, E., Andres, B., Savchynskyy, B., and Rother, C. (2016). InstanceCut: from Edges to Instances with MultiCut. *arXiv preprint arXiv:1611.08272*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). SSD: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully Convolutional Models for Semantic Segmentation. In *Conference on Computer Vision and Pattern Recognition*.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

Romera-Paredes, B. and Torr, P. H. S. (2016). Recurrent instance segmentation. In *arXiv:1511.08250*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Schwartz, E. L. (1977). Spatial mapping in the primate sensory projection: analytic structure and relevance to perception. *Biological cybernetics*, 25(4):181–194.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Ude, A., Atkeson, C. G., and Cheng, G. (2003). Combining peripheral and foveal humanoid vision to detect, pursue, recognize and act. In *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, volume 3, pages 2173–2178. IEEE.

Uhrig, J., Cordts, M., Franke, U., and Brox, T. (2016). Pixel-level encoding and depth layering for instance-level semantic segmentation. In *German Conference on Pattern Recognition*.

van den Brand, J., Ochs, M., and Mester, R. (2016). Instance-level Segmentation of Vehicles using Deep Contours. In *Asian Conference on Computer Vision*.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Wang, Z. and Bovik, A. C. (2006). Foveated image and video coding. *Digital Video, Image Quality and Perceptual Coding*, pages 431–457.

Zhang, Z., A., S., S., F., and R., U. (2015). Monocular object instance segmentation and depth ordering with CNNs. In *International Conference on Computer Vision*.

Zhang, Z., Fidler, S., and Urtasun, R. (2016a). Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *Conference on Computer Vision and Pattern Recognition*, pages 669–677.

Zhang, Z., S., F., and R., U. (2016b). Instance-level segmentation with deep densely connected MRFs. In *Conference on Computer Vision and Pattern Recognition*.