Thinking Outside the Box: Spatial Anticipation of Semantic Categories

Martin Garbade
garbade@iai.uni-bonn.de
Juergen Gall
gall@iai.uni-bonn.de

Computer Vision Group
University of Bonn
Germany

Abstract

For certain applications like autonomous systems it is insufficient to interpret only the observed data. Instead, objects or other semantic categories, which are close but outside the field of view, need to be anticipated as well. In this work, we propose an approach for anticipating the semantic categories that surround the scene captured by a camera sensor. This task goes beyond current semantic labeling tasks since it requires to extrapolate a given semantic segmentation. Using the challenging Cityscapes dataset, we demonstrate how current deep learning architectures are able to learn this extrapolation from data. Moreover, we introduce a new loss function that prioritizes on predicting multiple labels that are likely to occur in the near surrounding of an image.

1 Introduction

One of the core capabilities of humans intelligence is to make predictions about the environment. Humans are able to predict how the world around them will evolve in the near future and how their actions will affect it. Even without observing an entire scene, they can anticipate objects or surfaces that are close. This ability allows them to plan ahead and to efficiently interact with the world. Similar anticipation capabilities are also required for autonomous systems. For instance, the presence of semantic categories like pedestrians, bicyclists, cars, roads or sidewalks in the near surrounding of an autonomous vehicle has implications for the driving policy and safety measurements. These object categories, however, are not always within the field of view of the sensors attached to the vehicle and therefore need to be anticipated.

In this work, we propose the first approach that anticipates semantic categories outside the field of view of a camera. In order to evaluate this task, we propose a novel protocol for the large-scale Cityscapes dataset [3], which is the state-of-the-art benchmark for semantic urban scene understanding. In contrast to semantic image segmentation, which requires to infer the labels for each observed pixel, anticipation of semantic categories outside the field of view requires to infer the semantic labels in regions that are not observed. The anticipation task is not only more difficult due to missing data, it is also inherently non-deterministic since many solutions could be plausible. Since the true distribution of all plausible solutions for

a single image is unknown, we propose an evaluation metric that does not require a pixel-wise prediction but measures if the occurrence of a semantic class within a predefined region outside the image is correctly predicted.

Since the proposed task has not been addressed before, we introduce a baseline that infers the pixel-wise semantic labels in the observed region and the unobserved region outside the image. The baseline builds on a state-of-the-art convolutional neural network for image segmentation [2]. In addition, we propose a novel approach that consists of two networks. While the first network infers semantic labels for each observed pixel, the second network gradually anticipates the semantic categories outside the field of view from the previous output. For the second network, two different loss functions are investigated. We evaluate the proposed approach on the Cityscapes dataset [3] using the new protocol for spatial anticipation of semantic categories. The experimental evaluation shows that the proposed approach improves the baseline by a large margin.

2 Related Work

Since neural networks achieve impressive results in the domain of image classification [5, 8, 13, 14], they have also been successfully applied in the context of semantic image segmentation, *e.g.* in [1, 2, 6, 10]. The task that we address, however, has not been previously studied. The most related network architecture for semantic image segmentation is the approach by Chen *et al.* [2]. It is based on the ResNet architecture [5] and uses a couple of adaptations to make it suitable for semantic image segmentation. The main adaptation was the introduction of atrous-convolutions, which are convolutions with increased kernel sizes but with the same amount of parameters. By varying the kernel size of the atrous-convolution one can compute responses from the feature maps at different spatial resolutions which allows to control the size of the receptive field. In addition, the predictions are refined by a conditional random field [7].

Hallucinating semantics has been addressed in very few works. Liu *et al.* [9] developed an approach to reconstruct 3D scenes by simultaneously predicting depth and semantic labels from incomplete depth data. They propose a two-layer model representing both the visible and the hidden or occluded information. The approach also has some relations to in-painting methods like [11], which fill holes inside an image. In-painting methods, however, cannot be applied to anticipate semantics outside an image. For the task of proposal generation for an object detector, Ristin *et al.* [12] predict from large image patches potential bounding boxes that might contain objects of a relevant object category. While the bounding boxes can be outside the image patch, the approach aims at exploiting the local context within an image to reduce the inference time of an object detector. Using context for recognition tasks has also been extensively studied, for instance, in the seminal work by Torralba [15]. Recently, temporal anticipation has been studied by a few works. For instance, Vondrick *et al.* [16] predict feature representations for a video frame in the future, which can then be used to anticipate actions or objects that will occur next in a video.

3 Dataset for Anticipation of Semantic Categories

We propose the new task of spatial anticipation of semantic categories outside the field of view. The task requires to predict for a given image the categories that are most likely to

occur outside of it as illustrated in Figure 3. For evaluation, we introduce a new protocol for the Cityscapes dataset [3]. The Cityscapes dataset is recorded by an RGB camera mounted at the front of a car driving through urban scenes. We only use the images provided with fine-grained annotation. There are 2,975 images in the training set, 500 in the validation set and 1,525 images in the test set. For evaluation, we take the images from the validation set since the ground truth annotations for the test set are not publicly available. Following [3], we evaluate the performance on 19 classes ignoring the background class. The original images have the size of 1024×2048 pixels. For our task, we crop the validation images such that only the center region of 642×1282 pixels remains. The invisible region outside the cropped area is used to evaluate the anticipation performance.

For the evaluation, we report the accuracy for two evaluation criteria. The first evaluation criterion is a standard semantic image segmentation metric and compares the ground-truth segmentation map for the invisible region with the inferred pixel-wise semantic segmentation. It assumes that exactly one label is predicted for each pixel outside the cropping area of the original images. As for standard semantic image segmentation, we use the Jaccard index, also referred to as intersection over union (IoU), to measure the quality of the prediction. This evaluation approach has the weakness that it assumes that the ground-truth is deterministic and can be predicted at a pixel level. However, not even humans will be able anticipate semantic classes with such an accuracy. Moreover, an exact localization of the anticipated labels is unnecessary in a practical context. To account for this fact, we introduce an alternative evaluation metric. The unobserved area is subdivided into a grid of cells as shown in Figure 2. All labels that occur in the same cell are collected. If a label occurs in the same cell in both the ground-truth and the prediction map, it is counted as a true positive. Labels only occurring in the prediction map are false positives and labels exclusively occurring in the ground-truth map are considered as false negatives. We sum the true positives, false positives, true negatives, and false negatives over all cells and images for each class and compute the F_1 score, which is defined as the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{1}$$

The scores are then averaged over all classes.

4 SASNet: Convolutional Neural Network for Spatial Anticipation of Semantic Categories

To explore the task of spatial anticipation of semantic categories outside the field of view, we propose an approach that uses a convolutional neural network architecture as it is common for state-of-the-art approaches for semantic image segmentation. While the intermediate layers are based on the ResNet 101 structure [5] as in [2], the last layers and loss function differ from convolutional networks for image segmentation. Due to the task, the model also needs to be trained in a different way.

Figure 1 gives an overview of the proposed SASNet and the training procedure. The network is trained by providing masks for the visible and invisible regions for each training image. The mask divides the original image, which has a resolution of 1024×2048 pixels in our dataset, into an inner region Ω_1 of 642×1282 pixels and an outer region Ω_2 that is set to zero as shown in Figure 1 a). We then sample random crops from the images as shown

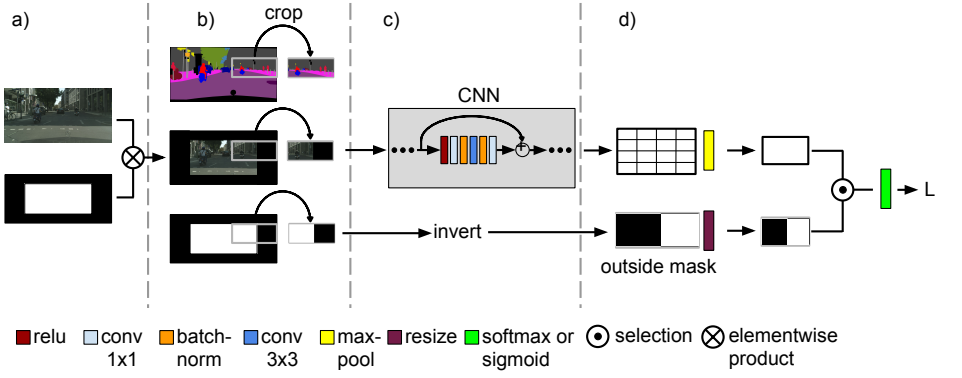

Figure 1: Training procedure for SASNet. The SASNet is trained on crops of masked images. The mask marks the visible and the invisible regions. A detailed description is given in Section 4.

in Figure 1 b). The crops of size 321×321 pixels are taken from the image, its ground-truth segmentation mask, and the visibility mask. Note that the ratio of the visible and invisible area varies among the crops. The random crops are our training set $\mathcal{T}$.

Figure 1 c) illustrates the first part of the network. As base architecture for the convolutional neural network, we choose the DeepLab model [2] based on the ResNet 101 structure [5]. We omit the conditional random field as well as the loss layer and instead process the unnormalized network output y to compute the loss for the unobserved region Ω_2. We investigate two different loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$. The first loss $\mathcal{L}_1$ is given by the softmax cross entropy:

$$\mathcal{L}_1 = -\sum_{t\in\mathcal{T}}\sum_{i\in\Omega_2^t}\sum_{c\in\mathcal{C}} \hat{y}_{ic} \log\left(\frac{e^{y_{ic}}}{\sum_{c'\in\mathcal{C}} e^{y_{ic'}}}\right) \tag{2}$$

where $\hat{y}_{ic}$ is the class probability of the ground truth label of pixel i, which is one for the true class and zero otherwise. y_{ic} denotes the unnormalized predictions of the network for pixel i and class c.

The second loss $\mathcal{L}_2$ measures the anticipation error in accordance with the proposed second evaluation criterion described in Section 3, *i.e.* only the classes occurring in each cell in the region Ω_2 should be predicted. This can be efficiently realized as illustrated in Figure 1 d) by adding a max pooling layer with kernel size and stride k:

$$\tilde{y}_{i_k c} = \max_{\Delta i\in\mathcal{N}_{i_k}} \left\{y_{i_k+\Delta i,c}\right\} \tag{3}$$

where $\mathcal{N}_{i_k}$ is the $k \times k$ neighborhood of pixel i_k, *i.e.* $\tilde{y}_{i_k c}$ is the maximum value for each class c in each cell i_k. It is important to note that the kernel size does not need to be equal to the cell size used for evaluation as we will show in Section 5.1. Due to the max pooling, Ω_2 has been reduced to the number of cells $\Omega_{k,2}$. We therefore also resize the mask to $\Omega_{k,2}$. For the cells of the invisible region, we compute the second loss $\mathcal{L}_2$ using the sigmoid cross entropy:

$$\mathcal{L}_2 = -\sum_{t\in\mathcal{T}}\sum_{i_k\in\Omega_{k,2}^t}\sum_{c\in\mathcal{C}} \hat{y}_{i_k c} \log\left(\frac{1}{1+e^{-\tilde{y}_{i_k c}}}\right) \tag{4}$$

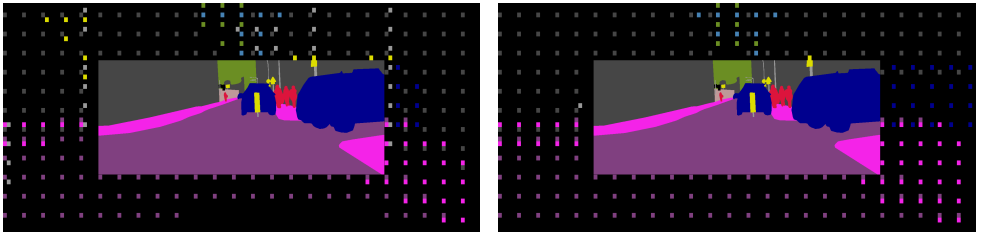

Figure 2: The F_1 score is computed for cells outside the visible region and measures if for each cell the same labels are predicted (right) as they occur in the ground-truth segmentation map (left).

where $\hat{y}_{i_k c}$ is one if the cell i_k in the invisible region Ω_2^t of random crop $t \in \mathcal{T}$ contains the label of class $c \in \mathcal{C}$ and it is zero otherwise.

For inference, the network processes an image with binary mask, which is one for the image pixels (Ω_1) and zero for the regions where the semantic categories should be anticipated (Ω_2). For the first loss function $\mathcal{L}_1$, the network predicts for each pixel i the semantic label given by $\operatorname{argmax}_c \frac{e^{y_{ic}}}{\sum_{c'} e^{y_{ic'}}}$. For the second loss function $\mathcal{L}_2$, the network predicts for each cell i_k the set of labels

$$\left\{ c \in \mathcal{C} : \frac{1}{1+e^{-\tilde{y}_{i_k c}}} \geq 0.5 \right\}. \tag{5}$$

Figure 2 shows an example of such a prediction.

In Section 5.1, we show that SASNet performs better when we first perform standard semantic image segmentation on the visible region Ω_1 and then use the inferred labels as input for SASNet instead of the RGB values of the image. The accuracy can be further improved by performing the anticipation in successive steps where the region Ω_2 outside the image is gradually increased and the intermediate results are used as input for the next step as shown in the last row of Figure 3.

5 Experimental Evaluation

5.1 Implementation and Evaluation Details

We augment the training data by random scaling between 0.5 and 1.5, as well as random mirroring and random cropping as in [2]. The batch size is set to 10 and the learning rate is set to $2.5 \cdot 10^{-4}$. The learning rate of the batch normalization layer parameters are set to zero. This has shown to stabilize the training process [2]. The number of training iterations is 20,000. The training takes about 15 hours.

As described in Section 3, we report intersection over union (IoU) and the F_1 score computed for four different cell sizes. As cell size c, we choose 16×16, 24×24, 40×40 and 80×80 pixels with respect to the original resolution of the Cityscapes images. Both measures are only evaluated on the unobserved region Ω_2.

We evaluate the two loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$ discussed in Section 4. For $\mathcal{L}_2$, we have to define the kernel size k. In our experiments, we evaluate $\mathcal{L}_2$ with the four different kernel sizes 2×2, 3×3, 5×5 and 10×10. Since the previous layers of the network reduce the size

of the input image by factor 8, this corresponds to the cell sizes 16×16, 24×24, 40×40 and 80×80 pixels with respect to the input resolution.

As mentioned in Section 4, SASNet can be used to anticipate semantic categories outside the field of view from the raw RGB image data or from a pre-segmentation of the visible region Ω_1. We evaluate both cases and use [2] for image segmentation in the latter case. We denote the first version by color-SASNet and the second version by label-SASNet. In addition, the anticipation can be performed gradually. Depending on the number of steps, we subdivide Ω_2 into either 2, 3 or 4 enclosing regions as can be seen in Figure 3. For each step, we use the prediction of the previous step as input and anticipate the semantic categories for the next enclosing region until Ω_2 is fully covered. For initialization, we use the inferred semantic segmentation of the visible region Ω_1. [1]

5.2 Results

The quantitative results for the dataset described in Section 3 are summarized in Table 1. The first six rows compare the two loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$ if SASNet anticipates the semantic categories from the RGB image (color-SASNet). For both, the intersection over union (IoU) and the F_1 accuracy, $\mathcal{L}_1$ performs better than $\mathcal{L}_2$. We will, however, observe that this changes if the anticipation is performed gradually.

In all cases, the F_1 score increases for larger c values since this increases the cell size, which requires a lower localization accuracy. If we compare different values of k for $\mathcal{L}_2$, we observe that IoU is slightly higher for $k = 1 \times 1$ since a smaller k enforces the network to learn a better localization of the categories. The setting with $k = 1 \times 1$ is also the best for all c values of the F_1 score.

We now compare the difference of having one network (color-SASNet) or two networks (label-SASNet), one for semantic image segmentation and one for spatial anticipation. A qualitative comparison is also shown in Figure 3. If we compare the $\mathcal{L}_1$ loss, IoU increases from 26.1 to 30.7. For the $\mathcal{L}_2$ loss with $k = 1 \times 1$, IoU increases from 22.0 to 26.6. The F_1 scores also increase for both $\mathcal{L}_1$ and $\mathcal{L}_2$ by about 4 to 5% for all c values, except for $\mathcal{L}_1$ in the case of $c = 80$. We can conclude that label-SASNet outperforms color-SASNet.

As illustrated in Figure 3, the anticipation accuracy decreases if the distance to the visible image border becomes large. The anticipation can therefore performed gradually where the region Ω_2 grows in each step as described in Section 4. The quantitative results using 2, 3, or 4 steps are reported in Table 1. We first compare the impact of the number of steps for label-SASNet with $\mathcal{L}_1$ loss. The IoU increases from 30.7 to 33.5 if anticipation is performed in two steps. Further steps increase the accuracy only slightly. The F_1 scores are also slightly improved by estimating the semantic categories outside the image region gradually. If the $\mathcal{L}_2$ loss is used, we observe a large improvement for all values of k. The best results are achieved with two steps. For $k = 5 \times 5$, the F_1 scores increase from 31.4, 31.8, 34.6, 36.3 to 42.7, 43.5, 44.9, 45.3, for $c = 16$, 24, 40, 80 respectively. The IoU also increases from 26.6 to 35.0 for $k = 1 \times 1$. It actually even achieves a higher IoU than the best setting with $\mathcal{L}_1$ loss (33.9).

Since label-SASNet learns to extrapolate a semantic segmentation of a scene, we also compare it to a standard approach for extrapolation. Using the inferred semantic segmentation of the visible region Ω_1, we replicate the labels of the border. The results are shown in

[1] The scripts, source code, and models used for evaluation are publicly available at https://pages.iai.uni-bonn.de/gall_juergen/projects/spatial_anticipation/spatial_anticipation.html.

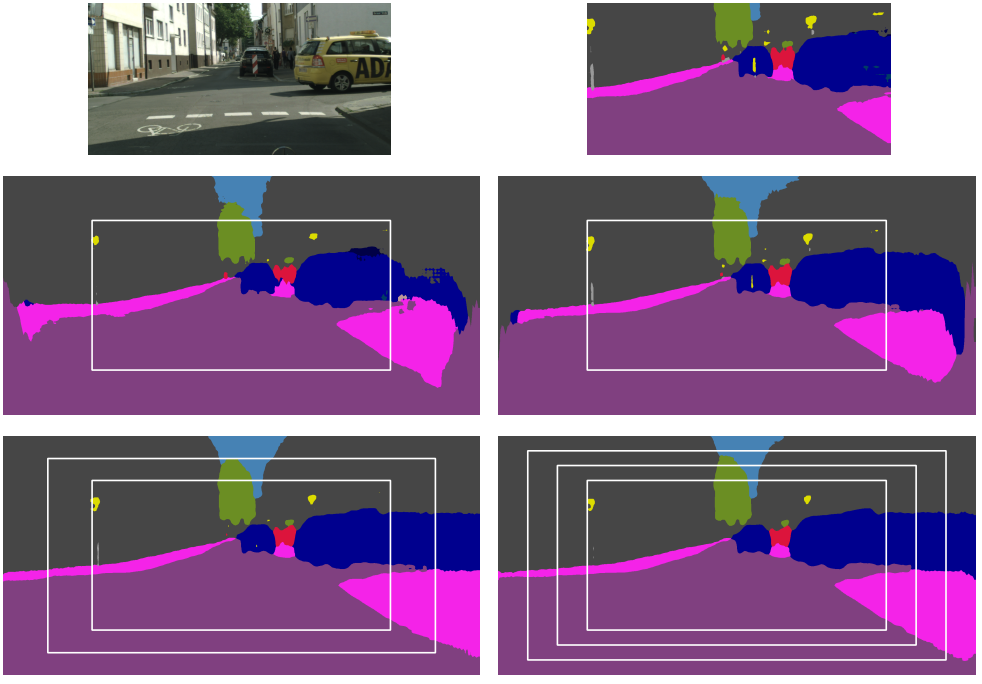

Figure 3: Qualitative results for the pixel-wise label prediction using $\mathcal{L}_1$ loss. The first row shows an RGB image and the inferred semantic segmentation using [0]. The second row shows the result for color-SASNet (left), which uses the RGB image of the first row as input, and for label-SASNet (right), which uses the inferred labels as input. The inner white rectangle marks the boundary between observed and unobserved regions Ω_1 and Ω_2. The label-SASNet anticipates the semantic labels in Ω_2 better than color-SASNet. The last row shows the result of label-SASNet if the prediction is performed in two (left) or three steps (right). The additional white rectangles mark the growing regions that are predicted in each step. Compared to the second row, the labels are better anticipated at the border.

the last row of Table 1. The border replication achieves 34.7 IoU but it performs poorly for the F_1 measure. This shows that the IoU measure, which is dominated by semantic classes that cover many pixels, is less appropriate than the F_1 measure for evaluating spatial anticipation.

We can conclude that anticipating semantic categories with two steps improves the accuracy by a large margin. The proposed $\mathcal{L}_2$ loss performs better than the $\mathcal{L}_1$ loss with respect to the F_1 score as well as IoU. Although the impact of k is very low, $k = 1 \times 1$ is best if the accuracy is measured by IoU and $k = 5 \times 5$ works very well for any c value of the F_1 score. For the case of prediction in a single step, $\mathcal{L}_1$ performs better than $\mathcal{L}_2$ with respect to IoU and F1 score. However, as the number of iterations for the prediction and the size of the evaluation cell are increasing, $\mathcal{L}_2$ outperforms $\mathcal{L}_1$.

6 Conclusion

We have introduced a new task of anticipating semantic label information outside of an image. We investigated two evaluation metrics to assess the quality of the prediction. While the

Input	$\mathcal{L}$	steps	k	% IoU	c = 16	c = 24	c = 40	c = 80
RGB	$\mathcal{L}_1$			26.1	34.7	35.5	37.5	38.9
RGB	$\mathcal{L}_2$		1x1	22.0	26.6	27.2	29.6	32.7
RGB	$\mathcal{L}_2$		2x2	21.2	25.6	26.2	28.4	31.8
RGB	$\mathcal{L}_2$		3x3	21.2	25.8	26.4	28.5	32.0
RGB	$\mathcal{L}_2$		5x5	21.4	26.4	26.9	29.0	32.4
RGB	$\mathcal{L}_2$		10x10	20.3	25.2	25.6	27.5	30.7
Label	$\mathcal{L}_1$			30.7	39.7	40.2	41.7	38.0
Label	$\mathcal{L}_1$	2		33.5	41.3	42.3	43.2	44.0
Label	$\mathcal{L}_1$	3		33.9	42.0	42.7	43.3	43.4
Label	$\mathcal{L}_1$	4		33.9	42.3	43.0	43.6	43.9
Label	$\mathcal{L}_2$		1x1	26.6	30.8	31.4	34.0	35.3
Label	$\mathcal{L}_2$		2x2	26.3	30.3	30.9	33.4	35.3
Label	$\mathcal{L}_2$		3x3	26.2	31.0	31.5	33.9	35.8
Label	$\mathcal{L}_2$		5x5	26.7	31.4	31.8	34.6	36.3
Label	$\mathcal{L}_2$		10x10	26.6	30.5	31.1	34.1	36.3
Label	$\mathcal{L}_2$	2	1x1	**35.0**	42.3	43.0	44.1	43.7
Label	$\mathcal{L}_2$	2	2x2	34.8	42.6	**43.5**	44.7	45.1
Label	$\mathcal{L}_2$	2	3x3	34.8	42.5	43.4	44.6	44.9
Label	$\mathcal{L}_2$	2	5x5	34.6	**42.7**	**43.5**	**44.9**	45.3
Label	$\mathcal{L}_2$	2	10x10	34.6	42.3	43.3	44.6	**45.4**
Label	$\mathcal{L}_2$	3	1x1	33.9	41.2	41.9	43.1	43.1
Label	$\mathcal{L}_2$	3	2x2	33.7	41.4	42.1	43.6	44.1
Label	$\mathcal{L}_2$	3	3x3	33.7	41.6	42.1	43.5	44.0
Label	$\mathcal{L}_2$	3	5x5	33.7	41.7	42.3	43.8	44.3
Label	$\mathcal{L}_2$	3	10x10	33.7	41.2	41.8	43.3	44.1
Label	$\mathcal{L}_2$	4	1x1	32.8	40.1	40.6	42.0	42.5
Label	$\mathcal{L}_2$	4	2x2	32.8	40.1	40.7	42.4	42.8
Label	$\mathcal{L}_2$	4	3x3	32.7	40.4	40.9	42.5	43.0
Label	$\mathcal{L}_2$	4	5x5	32.7	40.4	40.9	42.7	43.0
Label	$\mathcal{L}_2$	4	10x10	32.8	39.8	40.3	42.3	43.4
Label	Extrapolation			34.7	11.6	12.8	14.0	16.8

Table 1: Quantitative results for spatial anticipation on the Cityscapes dataset [3]. RGB or Label denote if color-SASNet or label-SASNet are used. $\mathcal{L}_1$ stands for pixel-wise loss and $\mathcal{L}_2$ for the cell-wise loss. k is the kernel size and stride used to compute the $\mathcal{L}_2$ loss during training. The third column indicates if the SASNet was applied gradually using 2, 3 or 4 steps. The fifth column is the pixel-wise evaluation using % IoU. The other columns are F1 scores expressed in % computed for the cell-wise evaluation. The size of the cells is specified as c. The last row shows the result if the labels are extrapolated by label replication.

first metric measures pixel-wise accuracy, the second metric relaxes the required localization accuracy and requires only the prediction of categories occurring in cells. In addition, we have proposed a neural network for spatial anticipation and investigated two different loss functions. From our experimental evaluation, we conclude that the most effective configuration uses two networks. The first one infers a pixel-wise segmentation within the visible area and the second one anticipates categories outside of the image from the segmented image. If the second network is applied gradually, the anticipation accuracy increases by a large margin. In this configuration, training the second network using the cell-wise loss performs for both evaluation metrics better than a pixel-wise loss. For the pixel-wise metric, it is most effective to choose the smallest possible kernel size for the loss function. For the cell-wise metric, the kernel size $k = 5 \times 5$ has shown to perform very well for any cell size used for evaluation. Although we have demonstrated the anticipation capabilities of the proposed approach, more research is required to achieve human performance. The proposed protocol and evaluation measure will facilitate the research on spatial anticipation.

Acknowledgements

The work has been financially supported by the DFG project GA 1927/2-2 as part of the DFG Research Unit FOR 1505 Mapping on Demand (MoD).

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *International Conference on Learning Representations*, 2015.

[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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[4] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning Hierarchical Features for Scene Labeling. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 35, pages 1915–1929, 2013.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[6] Alexander Kolesnikov and Christoph H Lampert. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation. In *European Conference on Computer Vision*, pages 695–711, 2016.

[7] Philipp Krähenbühl and Vladlen Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *Advances in Neural Information Processing Systems*, pages 109–117, 2011.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[9] Miaomiao Liu, Xuming He, and Mathieu Salzmann. Building Scene Models by Completing and Hallucinating Depth and Semantics. In *European Conference on Computer Vision*, pages 258–274, 2016.

[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[11] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context Encoders: Feature Learning by Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[12] Marko Ristin, Juergen Gall, and Luc Van Gool. Local Context Priors for Object Proposal Generation. In *Asian Conference on Computer Vision*, pages 57–70, 2013.

[13] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015.

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[15] Antonio Torralba. Contextual Priming for Object Detection. In *International Journal of Computer Vision*, volume 53, pages 169–191, 2003.

[16] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating Visual Representations from Unlabeled Video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–106, 2016.