

A new metric for evaluating semantic segmentation: leveraging global and contour accuracy

Eduardo Fernandez-Moral¹, Renato Martins¹, Denis Wolf², and Patrick Rives¹

Abstract—Semantic segmentation of images is an important issue for intelligent vehicles and mobile robotics because it offers basic information which can be used for complex reasoning and safe navigation. Different solutions have been proposed for this problem along the last two decades, where recent deep neural networks approaches have shown very promising results in the context of urban navigation. One of the main problems when comparing different semantic segmentation solutions is how to select an appropriate metric to evaluate their accuracy. On the one hand, classic metrics do not measure properly the accuracy on the object contours, which is important in urban driving to differentiate road from sidewalk for instance. On the other hand, contour-based metrics [1] disregard the information far from class contours. This paper explores the problem multi-modal image segmentation, and presents a new metric to leverage global and contour accuracy in a simple formulation. This metric is validated with the evaluation of several semantic segmentation solutions that exploit RGB-D images to rank these solutions taking into account the quality of the segmented contours. We also present a comparative analysis of several commonly used metrics together with a statistical analysis of their correlation.

I. INTRODUCTION

The problem of semantic segmentation consists of associating a class label to each pixel of a given image, resulting in another image of semantic labels. This problem of image interpretation is highly relevant in the context of mobile robotics and autonomous vehicles, for which accurate information of the objects in the scene may be applied for decision making or safe and robust navigation among others [2]. Semantic segmentation has seen a rapid progress over the past decade. Recent advances achieved by training different types of Convolutional Neural Networks (CNN) have improved notably the accuracy of state-of-the-art techniques [3], [4], [5], [6], [7], [8], [9]. Among the many CNN architectures available, convolutional encoder-decoder networks are particularly well adapted to the problem of pixel labeling. The encoder part of the network creates a rich feature map representing the image content and the decoder transforms the feature map into a map of class probabilities for every pixel of the input image. Such operation takes into account the pooling indices to upsample low resolution features into the original image resolution. The advantages of this class of network were presented in [6], [7]. The approach in [7] was later extended to a Bayesian framework in [8] to provide the probabilities associated to the pixel labels. Apart from end-to-end CNNs, Conditional Random Fields (CRFs)

have also been used for scene semantic segmentation [10], [4], [11]. In [12], a CNN model is used to extract features which are feed to a Support Vector Machine-based CRF to increase the accuracy of image segmentation.

The recent availability of 3D range sensors and RGB-D cameras has also been exploited for semantic segmentation [13], [3], [14], [9]. An initial exploration of adding geometric information besides color (e.g., depth images) was addressed in [13], but the global accuracy improvement was marginal. Later, [3] presented an approach where depth information is encoded into images containing horizontal disparity, height above the ground and angle with gravity, which outperforms previous solutions using raw depth for indoor scenes. On the other hand, [9] proposes to fuse depth features and color features in the encoder part of an encoder-decoder network. A CNN-based approach for joint pixel-wise prediction of semantic labels, depth and surface normals was presented in [15].

The appearance of public datasets and benchmarks for semantic segmentation in urban environments, both from virtual and real scenarios [16], [17], [18], facilitates the comparison of solutions, and promotes the standardization of comparison metrics. Still, the choice of the most appropriate metrics to evaluate semantic segmentation is a problem itself, which gains relevance with the increase of performance and complexity of semantic segmentation techniques.

In this paper, we investigate the problem of finding a single accuracy metric that accounts for both global pixel classification and good contour segmentation, that we use to compare the accuracy of different CNN architectures and different combinations of visual and range data. We propose a new metric based on [19] and [1] which makes use of the Jaccard index to account for boundary points with a candidate match belonging to the same class in the target image. As we show in our experiments, this metric blends the characteristics of the Jaccard index (which is the *de facto* standard in semantic segmentation) and the border metric BF in a simple formulation, thus allowing to compare easily the outputs of different segmentation solutions.

II. BACKGROUND

A. Semantic segmentation from color and depth

The combination of RGB and depth information to produce RGB-D images is interesting for many applications. Using color and depth information has also proven to be useful for semantic segmentation [3], [14], [9]. For instance, range information can be exploited for inferring discontinuities among different object classes. Such discontinuities

¹INRIA Sophia Antipolis - Méditerranée, 2004 Route des Lucioles - BP 93, 06902 Sophia Antipolis, France. e-mail: name.surname@inria.fr

²University of Sao Paulo - ICMC/USP, Brazil. denis@icmc.usp.br

are highly relevant in the context of urban driving, as they can help greatly to segment the boundaries of the road and sidewalk, and can help to identify dynamic objects from the scene. However, it's not clear yet how these two types of data should be fed into the CNN, and which network architecture is optimal for the problem. We evaluate some state-of-the-art network models and different preprocessing of the RGB-D information in order to find out which combination of network and input data is better suited for semantic segmentation.

B. Semantic segmentation metrics

Comparing the accuracy in segmentation approaches is commonly carried out through different global and class-wise statistics such as global precision, class-wise precision, confusion matrix, F-measure or the Jaccard index (also called "Intersection over Union" [IoU]). Global metrics like the precision are only an acceptable indicator to evaluate different solutions when the different semantic categories have a similar relevance (both in terms of frequency of appearance and practical importance). But this is not the case in most applications, where objects which have fewer pixels may be significantly more relevant than others (e.g., "traffic light" or "cyclist" classes versus the "sky" in the context of autonomous vehicles). On the other hand, class-wise metrics (e.g., [7], [9]) avoid the previous limitation, but computing accuracies for each class individually means that we cannot compare different segmentation solutions directly (without specifying quantitatively the relevance of each class). An alternative metric is to average the chosen class-wise metric m according to the total number of classes n (e.g., $\bar{m} = \sum_{i=1}^n m_i/n$). This class-wise average is less affected by imbalanced class frequencies than global metrics, and are the reference to rank different solutions in semantic segmentation benchmarks¹.

Another relevant aspect when evaluating segmentation approaches is to measure the quality of the segmented contours. [20] proposes to measure the ratio between correct and wrong classified pixels in a region surrounding the class boundaries, instead of considering all image pixels. Other contour-based metrics include the Berkeley contour matching score [19], the boundary-based evaluation [21] and the contour-based score [1]. All these measures are based on the matching between the class boundaries in the ground truth and the segmented images. [21] computes the mean and standard deviation of a boundary distance distribution between pairs of boundary images. [19] computes the F1-measure from precision and recall values using a distance error tolerance θ to decide whether a boundary point has a match or not. [1] proposes the BF score as an adaptation of [19] to multi-class segmentation, computed as the average of F_1 scores over the classes present in the segmented image.

The trade-off between global and contour segmentation is an important issue since both: a high rate of correctly labeled pixels and a good contour segmentation are desirable.

TABLE I: Class confusion matrix and notation.

		Predicted class	
		Positive	Negative
True class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

In [1], the authors suggest to use both the Jaccard index and BF as accuracy metrics to capture different aspects of the segmentation quality (global and contour). However, when the problem consists of ranking different segmentation approaches a single measure is required to compare directly different results. This problem is highly relevant when exploring different neural network architectures for semantic segmentation as we do in this paper. Besides, accuracy metrics which are also influenced by the quality of boundaries are interesting as loss functions to train segmentation models with machine learning.

C. Standard accuracy metrics

This section describes the most common metrics used for semantic segmentation, where the notation is given in table I. A general analysis of accuracy metrics for classification tasks can be found in [22].

The "accuracy" is the ratio of the correctly classified elements over all available elements:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

The "precision", or positive predictive value (PPV), is the relation between true positives and all positive predictions:

$$Precision = \frac{TP}{TP + FP}. \quad (2)$$

The "Recall", or true positive value (TPV), is the relation between true positives and all positive elements:

$$Recall = \frac{TP}{TP + FN}. \quad (3)$$

The F-measure [23] is a widely used metric to evaluate classification results, which consists of the harmonic mean of precision (2) and recall (3) metrics:

$$F_\beta = \frac{(\beta^2 + 1)TP}{(\beta^2 + 1)TP + \beta^2 FN + FP} \quad (4)$$

where β is scaling between the precision and recall. Considering $\beta = 1$, leads to the widely used F1-measure:

$$F_1 = \frac{2TP}{2TP + FN + FP}. \quad (5)$$

Another common metric to evaluate the results of classification is the Jaccard index (JI), or Intersection-over-Union:

$$JI = \frac{TP}{TP + FN + FP}. \quad (6)$$

Global accuracy metrics are not appropriate when class frequencies are unbalanced, which is the case in most scenarios both in real indoor and outdoor scenes, since they are biased by the dominant classes. To avoid this, the metrics above are usually evaluated per-class, and their result is averaged over the total amount of classes.

¹<https://www.cityscapes-dataset.com/benchmarks/>

The confusion matrix (C), is a squared matrix where each column represents the instances in a predicted class while each row represents the instances in an actual class, where a value C_{ij} represents the elements of the class i which are classified as the class j :

$$C_{ij} = |S_{gt}^i \circ S_{ps}^j| \quad (7)$$

with S_{gt}^i and S_{ps}^j being the binarized maps of the ground truth class i and predicted class j respectively, (\circ) represents the element-wise product and ($|\cdot|$) is the L1 norm. Note that the confusion matrix is also useful to compute the above metrics in a class-wise manner, e.g.:

$$JI^k = \frac{C_{kk}}{\sum_{i=1}^n C_{ik} + \sum_{j=1}^n C_{kj} - C_{kk}}. \quad (8)$$

III. THE NEW METRIC BJ

This section describes a new metric for supervised segmentation which measures jointly the quality of the segmented regions and their boundaries. Our metric is inspired by the BF score presented in [1], which is defined as follows. Let's call B_{gt}^c the boundary of the binary map of the S_{gt}^c of class c in the ground truth and likewise, B_{ps}^c for its predicted segmentation. For a given distance threshold θ , the precision P^c and the recall R^c for class c are defined as:

$$P^c = \frac{1}{|B_{ps}^c|} \sum_{x \in B_{ps}^c} [[d(x, B_{gt}^c) < \theta]] \quad (9)$$

$$R^c = \frac{1}{|B_{gt}^c|} \sum_{x \in B_{gt}^c} [[d(x, B_{ps}^c) < \theta]] \quad (10)$$

with $[[\cdot]]$ the Iversons bracket notation, where $[[b]] = 1$ if $b = true$ and 0 otherwise, and $d(\cdot)$ the Euclidean distance measured in pixels. The F_1^c measure for class c is given by:

$$BF^c = F_1^c = \frac{2 \cdot P^c \cdot R^c}{P^c + R^c}. \quad (11)$$

The BF in (11) has two main drawbacks. First, it disregards the content of the segmentation beyond the threshold distance θ under which boundaries are matched. Second, the results of this metric depends on a discrete filtering of the distribution of boundary distances, so that the same score is obtained for different segmentations (with different perceptual quality) as far as the same amount of boundary pixels are within the distance θ . This is shown in table II, which shows different infra and over-segmentations with their corresponding scores.

To overcome these shortcomings we propose to compute the distances from the boundary binary map to the binary map of the predicted segmentation $B_{gt}^c \rightarrow S_{ps}^c$ for a given class c to obtain the amount of true positives ($TP_{B_{gt}}^c$) and false negatives (FN^c). Similarly, we compute the distance from the boundary of the predicted segmentation to the binary map of the ground truth $B_{ps}^c \rightarrow S_{gt}^c$ for class c to obtain the amount of true positives ($TP_{B_{ps}}^c$) and false positives (FP^c). The total number of true positives is defined as ($TP^c = TP_{B_{gt}}^c + TP_{B_{ps}}^c$).

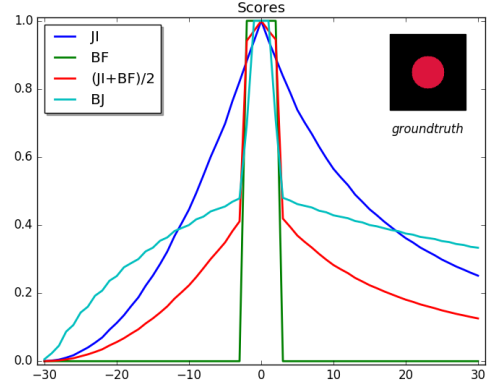


Fig. 1: Per-class scores of the segmented circle (top-right) for different levels of infra/over segmentation. The parameter θ is set to 4 pixels for both BF and BJ, which corresponds to 0.0075 of the image diagonal.

Note that while the BF measure is based on boundary-to-boundary matches, our proposed BJ score is boundary-to-object. To avoid the second shortcoming, we propose to measure the values above with a continuous measure of the boundary distances, so that the following values are defined:

$$TP_{B_{gt}}^c = \sum_{x \in B_{gt}^c} z \text{ with } z = \begin{cases} 1 - (d(x, S_{ps}^c)/\theta)^2 & \text{if } d(x, S_{ps}^c) < \theta \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

$$FN^c = |B_{gt}^c| - TP_{B_{gt}}^c \quad (13)$$

$$TP_{B_{ps}}^c = \sum_{x \in B_{ps}^c} z \text{ with } z = \begin{cases} 1 - (d(x, S_{gt}^c)/\theta)^2 & \text{if } d(x, S_{gt}^c) < \theta \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

$$FP^c = |B_{ps}^c| - TP_{B_{ps}}^c \quad (15)$$












Then, the score for class c , which we call *Boundary Jaccard* (BJ^c) is defined according to the Jaccard index:

$$BJ^c = \frac{TP^c}{TP^c + FP^c + FN^c}. \quad (16)$$

This new score is not zero when the ground truth and the predicted segmentation for a given class have some overlapping ($|S_{gt}^c \cup S_{ps}^c| > 0 \Rightarrow BJ^c > 0$). This behavior is similar for the metric JI^c but not for BF^c . On the other hand, BJ^c increases when the boundaries of ground truth and predicted segmentation get closer, like for BF^c , but with a more continuous behavior than the latter. Fig. 1 shows an example to illustrate the behavior of the metrics BJ^c , BF^c and JI^c for different levels of infra/over segmentation, as shown in table II.

Finally, in order to compute the per-image BJ score, we average the BJ^c scores over all the classes present either in the ground truth or in the predicted segmentation, as in [1]. This score is computed as the average of per-image BJ's over the number of images contained in the sequence.

TABLE II: Examples of infra-segmentation and over-segmentation of a pedestrian from the Cityscapes dataset. The ground truth corresponds to figure in the center.

										
0	.12	.45	.64	.86	$\leftarrow \mathbf{JI} \rightarrow$.88	.77	.66	.54	.30
0	0	0	0	.99	$\leftarrow \mathbf{BF} \rightarrow$.99	0	0	0	0
0	.20	.46	.47	.77	$\leftarrow \mathbf{BJ} \rightarrow$.79	.64	.50	.50	.48

IV. EXPERIMENTAL ANALYSIS OF ACCURACY METRICS

This section presents a number of qualitative and quantitative results showing the accuracy of different types of CNN trained and tested on the Virtual KITTI [17] and KITTI [18] datasets for the comparison of the different evaluation metrics. The results indicate that measuring accuracy in the neighborhood of class borders is useful to compare different solutions without the need to provide class weights. Furthermore, the proposed metric BJ is correlated with both JI and BF (i.e. it captures characteristics of these two scores). In the following experiments we focus our attention on the point of evaluating different accuracy metrics as it's the aim of this paper, rather than searching to improve the semantic segmentation results. Note that having an adequate metric is a prerequisite for the latter task.

A. CNN architectures for semantic segmentation from RGB-D data

Using color and depth information has proven to be useful for semantic segmentation [3], [14], [9]. However, it's not clear yet how these two types of data should be fed into the CNN, and which network architecture is optimal for the problem. Without trying to solve this problem, we just describe here several solutions in order to compare later the suitability of different accuracy metrics. The network models analyzed in the next section are FuseNet [9], SegNet [7], and some modified versions of the latter that we describe here. We introduce a modification of SegNet to obtain a more compact network which we call Compact Encoder Decoder Convolutional Neural Network (CEDCNN). This network model reduces the number of consecutive convolution filters (convolution+batch normalization+ReLU) to reduce the complexity and non-linearity of the model. This model is faster and easier to train, at the cost of being less descriptive and accurate. We also employ a modification of SegNet which is similar to [14], called SegNet2, with two separate networks for color and geometric information, whose result is concatenated and filtered by an additional convolution layer. In the same way as for SegNet2, we also modify the CEDCNN to obtain a new network called CEDCNN2, with two different pipelines to extract feature maps from color and geometric information separately.

B. Comparison of different metrics

We provide a quantitative analysis of the behavior of different metrics with infra-segmented and over-segmented objects, as shown in Table II. For that, we create synthetic segmentations of the ground truth of different object classes, e.g., “traffic sign” or “pedestrian”. For instance, using the “pedestrian” class shown in table II, we produce infra-segmented objects by removing layers of pixels from its boundary, such as the segmentations at the left of table II. Conversely, we produce over-segmented objects by adding layers of labeled pixels beyond the boundary, see the images at the right of table II. Fig. 1 shows the relation of the scores JI, BF and BJ with respect to the amount of infra/over-segmentation. The horizontal axis represents the amount of infra-segmentation (negative values) and over-segmentation (positive values) according to the number of 1-pixel layers removed or added to the ground truth, which is represented at the center of this graph, where all scores are 1. We also show the mean of JI and BF to indicate that a mix of these two is closer to the proposed BJ.

We observe that the JI has the most gradual behavior since it considers all pixels equally. The BF, which considers only the boundaries, shows a discontinuous trend on the threshold parameter used to distinguish boundary inliers from outliers. The JI and the BF are averaged to obtain a score that accounts for both: the number of pixels correctly labeled and the quality of contours of the segmentation. While the discontinuity of this averaged metric is less severe than for BF, it is still undesirable because the score depends highly on the threshold value θ . In contrast, our BJ score shows a continuous behavior because it does not rely on a discontinuous filter like BF. The BJ score is higher than JI for infra-segmented objects, which is interesting to avoid neglecting infra-segmentations. The BJ score is close to 0.5 for over-segmented objects with bad contour segmentation, which is reasonable since an over-segmentation is always preferable to a miss-classification. Besides, the effect of over-segmentations penalizes the BJ scores of the surrounding objects in the image, resulting in lower average BJ scores for images with larger over-segmentations.

C. Semantic segmentation on KITTI and Virtual KITTI

We use the Virtual KITTI dataset [17] for training and testing different models for semantic segmentation, which are later fine-tuned using the original KITTI dataset. The

Virtual Kitti dataset contains RGB, depth and labeled images with 13 classes: *sky, sidewalk, tree, vegetation, building, road, guard rail, traffic sign, traffic light, pole, car, van and truck*. It is composed of 5 virtual scenarios resembling those from the KITTI dataset [18], generated by simulating different illumination and weather conditions. Our training data is composed of 3846 observations chosen along different parts of the 5 scenarios contained in the dataset, scattering the selected images through the different sequences with different conditions (clone, fog, morning, overcast, rain and sunset). Each model is trained independently from scratch from the same training data. The test data used to produce the results shown in the following tables is composed of 1266 images selected from different sections of these datasets.

First, we evaluate different ways to feed geometric information into SegNet, which is trained from images of different types: color (RGB), raw depth (D) encoded in one channel with 16 bits for centimeter precision, normal vectors plus depth (ND), and normal vectors plus elevation from the ground (NE). The images ND and NE are encoded as 3-channel images with 8 bits per channel, with 2 channels containing the first two components of the normal directions and the third channel containing depth or elevation, accordingly scaled to 8 bits [3]. Table III presents the accuracy measured by different traditional scores, including global accuracy (GA) and different per-class metrics which are averaged per-image: the F1-measure, Jaccard index (JI), recall on class boundaries (TO), Jaccard index on boundaries (TJ), BF-measure and the boundary Jaccard measure (BJ) proposed here. All boundary measures are parametric, with a single parameter to set the width of the boundary region to consider (in pixels), as indicated in Table III. The first 5 rows of the table correspond to SegNet for different inputs. We observe that the combination of surface normal directions plus depth or elevation achieve the best results, with slightly better accuracy for normals plus depth (ND). These outperform the accuracy obtained using RGB, raw depth, and the case of 4-channel RGBD input which concatenates RGB with raw depth (with 8 bits for each color channel and 16 bits for depth)². Regarding the accuracy of the model SegNet2, the use of input data from RGB-ND achieves the best results, for which most accuracy metrics indicate that it is the best combination of model and input data.

We analyse next other network architectures like FuseNet [9], together with the models introduced above: SegNet2, CEDCNN and CEDCNN2. The lower part of Table III shows the evaluation of these methods. The results show that FuseNet, which was designed for semantic segmentation of indoor images from RGB-D data, achieves a performance comparable with SegNet. The authors of FuseNet [9] argue that the relevant geometric features can be learned from raw depth by the CNN without the need of previous transformations. However, we observe a relevant improvement by comparing the results of FuseNet using RGB-D vs. RGB-

ND, for which the surface directions contribute to improve the accuracy. This can be explained from the fact that the datasets of our experiments come from a forward facing camera mounted in a car and thus, the surface directions have some invariants such as the angle with gravity, that constitute a relevant source of information.

Regarding the different accuracy metrics, we observe that they do not agree for ranking the accuracy of different models and input data. This occurs because the accuracy of low frequency classes have a large variability even for similar models, which is reflected in our results. On the other hand, BJ presents a more stable behavior for similar models, where even little changes on its value seem to be a good indicator to choose the best model according to the visualization of the predicted segmentation (see the accompanying video).

D. Correlation of different metrics

We measure the correlation of the different metrics used in the previous experiment. For that, we compute the per-image score on the segmented test sequence of Virtual KITTI (RGB-ND) obtained with the model CEDCNN2, and measure the correlation of the different metrics for ranking the quality of each segmented image. We employ the Spearman's rank correlation (ρ), which is a nonparametric measure of rank correlation, defined as the Pearson correlation coefficient between the ranked variables. It measures the statistical dependence between the ranking of different accuracy metrics. For a sample of size n , with the n raw scores X_i, Y_i , the Spearman's rank correlation is defined as

$$\rho = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}} \quad (17)$$

where rg_X, rg_Y are the ranks of the score distributions X, Y .

Fig. 2 shows the ranking correlations of the different accuracy metrics, where we observe that the BJ score is correlated to both JI and BF, showing that they capture similar information in this particular experiment. This result is intuitively observed from the shape of the mean of the JI and BF scores in Fig. 1. Notice that the correlations with BJ are higher than the correlations among other pairs of scores.

V. CONCLUSIONS

This paper addresses the problem of measuring the accuracy of semantic segmentation of images, which is an essential aspect when comparing different segmentation approaches. We present a new metric to compute the Jaccard index considering the border regions of the different classes to leverage global and contour accuracy, which is suitable for unbalanced class frequencies. We also present results for several CNN models using two public datasets. As the experiments show, the proposed metric encodes jointly the rate of correctly labelled pixels, and how homeomorphic is the segmentation to the real object boundaries. This aspect is highly relevant to segment correctly the limits of the road, its lanes or the sidewalk for autonomous driving applications. Our results can also be applied to other contexts beyond urban driving, like indoor navigation. In our future research,

²Note that the virtual dataset has "perfect" geometry, which explains the high accuracy rates using only geometric information.

TABLE III: Average (per-class and per-image) accuracy metrics for different combinations of CNN architectures and input data using color and geometric information (in %).

Model \ Metric	GA	F1	IoU	TO3	TO5	TO7	TJ3	TJ5	TJ7	BF3	BF4	BF5	BJ3	BJ4	BJ5
SegNet (RGB)	82.0	40.1	34.3	62.8	68.4	71.5	24.8	28.0	29.7	33.6	36.8	39.7	40.7	42.8	44.4
SegNet (D)	87.2	43.0	37.5	71.6	76.2	78.7	30.1	32.6	34.0	41.2	43.7	45.9	45.3	47.1	48.5
SegNet (ND)	88.2	45.0	39.3	68.2	74.7	78.0	28.4	32.4	34.3	41.7	44.7	47.3	46.0	48.1	49.6
SegNet (NE)	88.1	46.3	40.4	68.1	74.5	77.8	28.8	32.9	35.1	42.5	45.7	48.6	47.4	49.7	51.4
SegNet (RGBD)	78.1	40.1	33.1	59.8	64.7	69.0	23.8	25.6	27.4	32.1	35.2	37.8	37.7	40.3	42.4
SegNet2 (RGB-D)	93.7	52.4	47.1	82.3	85.8	88.4	41.8	44.4	45.9	52.1	53.4	56.1	54.9	55.3	56.0
SegNet2 (RGB-ND)	94.1	52.6	47.3	82.0	85.7	87.6	41.2	43.9	45.1	52.4	54.0	56.3	55.2	55.9	56.7
SegNet2 (RGB-NE)	93.9	52.3	47.2	81.9	85.8	87.5	40.9	44.1	45.2	52.5	53.9	55.8	55.0	55.6	56.5
FuseNet (RGB-D)	90.7	50.3	45.1	81.4	85.2	87.9	41.4	44.0	45.2	51.4	52.5	54.1	52.9	54.0	54.5
FuseNet (RGB-ND)	91.9	52.0	46.2	81.2	85.0	87.6	41.3	43.7	45.3	51.8	53.0	55.1	53.3	54.7	55.0
CEDCNN (RGB)	90.3	47.6	42.1	75.8	80.8	83.2	34.4	37.5	39.2	45.7	48.1	50.1	49.2	51.0	52.3
CEDCNN (D)	86.0	47.2	41.4	74.4	78.2	80.1	35.9	38.0	39.2	46.2	48.5	50.4	50.3	51.9	53.0
CEDCNN (ND)	89.1	48.4	42.6	71.6	77.8	80.9	32.5	36.5	38.5	46.2	49.0	51.2	49.8	51.9	53.3
CEDCNN (NE)	88.8	49.0	43.0	71.2	77.5	80.6	32.4	36.8	38.9	46.8	49.7	51.9	50.6	52.8	54.2
CEDCNN2 (RGB-ND)	93.2	52.2	46.9	80.9	85.2	87.3	40.7	43.4	44.7	51.5	53.3	54.8	54.1	55.5	56.5

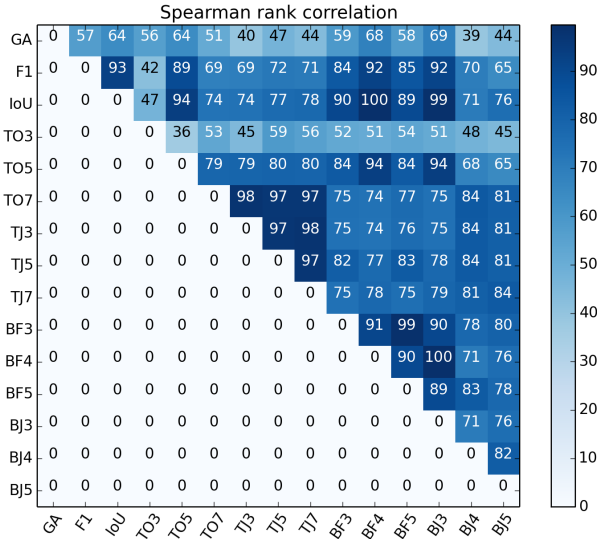


Fig. 2: Spearman's rank correlation of accuracy metrics.

we plan to study how to give more importance to the segmentation of contours during the CNN training phase by using our metric in the loss function.

REFERENCES

- [1] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?," in *BMVC*, 2013.
- [2] R. Drouilly, P. Rives, and B. Morisset, "Semantic representation for navigation in large-scale environments," in *ICRA*. IEEE, 2015.
- [3] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *ECCV*. Springer, 2014.
- [4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.
- [5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*. IEEE, 2015.
- [6] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *ICCV*. IEEE, 2015.
- [7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.
- [8] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.
- [9] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *ACCV*, 2016.
- [10] L. Ladický, P. Sturges, K. Alahari, C. Russell, and P. H. Torr, "What, where and how many? combining object detectors and crfs," in *European conference on computer vision*. Springer, 2010, pp. 424–437.
- [11] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *ICCV*. IEEE, 2015.
- [12] F. Liu, G. Lin, and C. Shen, "Crf learning with cnn features for image segmentation," *Pattern Recognition*, vol. 48, no. 10, pp. 2983–2992, 2015.
- [13] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.
- [14] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *IROS*. IEEE, 2015.
- [15] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *ICCV*. IEEE, 2015.
- [16] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*. IEEE, 2016.
- [17] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *CVPR*. IEEE, 2016.
- [18] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [19] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 26, no. 5, pp. 530–549, 2004.
- [20] P. Kohli, L. Ladický, and P. H. Torr, "Robust higher order potentials for enforcing label consistency," in *CVPR*. IEEE, 2008.
- [21] J. Freixenet, X. Muñoz, D. Raba, J. Martí, and X. Cufí, "Yet another survey on image segmentation: Region and boundary information integration," 2002.
- [22] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
- [23] C. J. Van Rijsbergen, *Information Retrieval*. Butterworths, 1979.