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

Eduardo Fernandez-Moral, Renato Martins, Denis Wolf, Patrick Rives

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

Eduardo Fernandez-Moral[1], Renato Martins[1], Denis Wolf[2], and Patrick Rives[1]

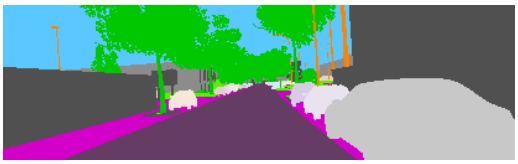
***Abstract*— Semantic segmentation of images is an important problem for mobile robotics and autonomous driving because it offers basic information which can be used for complex reasoning and safe navigation. Different solutions have been proposed for this problem along the last two decades, and a relevant increment on accuracy has been achieved recently with the application of deep neural networks for image segmentation. One of the main issues when comparing different neural networks architectures is how to select an appropriate metric to evaluate their accuracy. Furthermore, commonly employed evaluation metrics can display divergent outcomes, and thus it is not clear how to rank different image segmentation solutions. This paper proposes a new metric which accounts for both global and contour accuracy in a simple formulation to overcome the weaknesses of previous metrics. We show with several examples the suitability of our approach and present a comparative analysis of several commonly used metrics for semantic segmentation together with a statistical analysis of their correlation. Several network segmentation models are used for validation with virtual and real benchmark image sequences, showing that our metric captures information of the most commonly used metrics in a single scalar value.**

I. INTRODUCTION

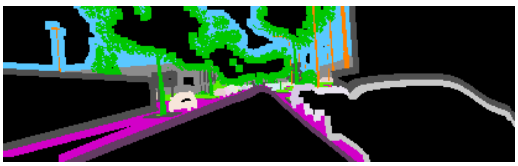
The problem of semantic segmentation consists of associating a class label to each pixel of a given image, resulting in another image of semantic labels, as shown in figs. 1a and 1b. This problem of image understanding is highly relevant in the context of mobile robotics and autonomous vehicles, for which accurate information of the objects in the scene may be applied for decision making or safe and robust navigation among others [1].

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 [2], [3], [4], [5], [6], [7], [8]. 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 [5], [6]. The approach in [6] was later extended to a Bayesian framework in [7] 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 [9], [3], [10]. In [11], 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.

(a) Colour image



(b) Annotated image of classes



(c) Image of class borders ($\theta = 5$)

Fig. 1: Extraction of class borders from an annotated image of labels from the Virtual KITTI dataset [12].

The recent availability of 3D range sensors and RGB-D cameras has also been exploited for semantic segmentation [13], [2], [14], [8]. 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, [2] 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. A different strategy for the same problem is presented in [8], which proposes to fuse depth features and color features in the encoder part of an encoder-decoder network. Another 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, both from virtual and real scenarios

[1]Lagadic team. INRIA Sophia Antipolis - Méditerranée. 2004 Route des Lucioles - BP 93, 06902 Sohpia Antipolis, France. Email: eduardo.fernandez-moral@inria.fr, renato-jose.martins@inria.fr, patrick.rives@inria.fr

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

[16], [12], [17], 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.

A. Contribution

In this paper, we investigate the problem of finding a single accuracy metric that accounts for both global pixel classification and good contour segmentation. We propose a new metric based on [18] and [19] 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.

B. Outline

The reminder of the paper is organized as follows. Section II-A reviews related works. In section II-B, we introduce the traditionally used segmentation evaluation metrics and their limitations. Section III describes our proposed metric. We present the different CNN architectures and the experimental results in section IV, considering simulated and real benchmark image sequences, such as the virtual KITTI and KITTI. Finally, in section VI, we draw conclusions and highlight future improvements and perspectives.

II. SEMANTIC SEGMENTATION METRICS

In this section, we review some recent related works and the background on commonly used evaluation metrics for semantic segmentation.

A. Related works

Comparing the accuracy of different semantic 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"). These metrics are described in more detail in section II-B. Global metrics like the precision may be a good 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., [6], [8]) 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., $\overline{m} = \sum_{i=1}^{n} m_i/n$). This class-wise average is less affected by imbalanced class frequencies than global metrics.

Another relevant aspect when evaluating segmentation approaches is to measure the quality of the segmentation around class 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 [18], the boundary-based evaluation [21] and the contour-based score [19]. 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. [18] 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. [19] proposes an adaptation of [18] to multi-class segmentation, where the score (BF) is 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. For instance, in the context of autonomous navigation, we are interested in segmenting accurately the borders of the road and sidewalk in order to delimit the navigable space for each agent. In [19], 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 in ranking different segmentation approaches based on their results, it is required to rely on a single measure so that different solutions can be directly compared. This problem is highly relevant, for instance, while using CNNs for semantic segmentation, because we are often interested in finding the set of hyperparameters which produce the best accuracy. This requires the comparison of multiple models using a single score. Besides, accuracy metrics which are also influenced by the quality of boundaries are interesting as loss functions to train the segmentation models.

B. Standard accuracy metrics

This section describes the most common metrics used for semantic segmentation. For reference, a general analysis of accuracy metrics for classification tasks can be found in [22].

The "accuracy", or the ratio of the correctly classified elements over all available elements can be calculated as follows:

$$Accuracy = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \tag{1}$$

whose notation is detailed in table I.

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

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{2}$$

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)

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

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{3}$$

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:

$$\mathrm{F}_\beta = \frac{(\beta^2+1)\mathrm{TP}}{(\beta^2+1)\mathrm{TP}+\beta^2\mathrm{FN}+\mathrm{FP}} \tag{4}$$

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

$$\mathrm{F}_1 = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FN}+\mathrm{FP}}. \tag{5}$$

Another common metric to evaluate the results of classification is the Jaccard index (JI):

$$\mathrm{JI} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}}. \tag{6}$$

Global accuracy metrics are not appropriate evaluation measures 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.

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. Thus, a value C_{ij} represents the elements of the class i which are classified as the class j:

$$C_{ij} = |S^i_{gt} \circ S^j_{ps}| \tag{7}$$

where S^i_{gt} and S^j_{ps} are 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.:

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

III. A NEW METRIC FOR SUPERVISED SEGMENTATION

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 [19], which is defined as follows. Let's call B^c_{gt} the boundary of the binary map of the S^c_{gt} of class c in the ground truth and likewise, B^c_{ps} for its predicted segmentation. For a given distance threshold θ, the precision for class c is defined as:

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

and the recall

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

with $[[\cdot]]$ the Iversons bracket notation, where $[[z]] = 1$ if z=true and 0 otherwise, and $d(\cdot)$ the Euclidean distance measured in pixels. The F^c_1 measure for class c is given by:

$$\mathrm{BF}^c = \mathrm{F}^c_1 = \frac{2 \cdot P^c \cdot R^c}{P^c + R^c}. \tag{11}$$

The BF in (11) has two main drawbacks. Firstly, it disregards the content of the segmentation beyond the threshold distance θ under which boundaries are matched. Secondly, 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.

In order to handle these shortcomings, we compute the distances from the boundary binary map to the binary map of the predicted segmentation $B^c_{gt} \to S^c_{ps}$ for a given class c to obtain the amount of true positives ($\mathrm{TP}^c_{B_{gt}}$) 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 $B^c_{ps} \to S^c_{gt}$ for class c to obtain the amount of true positives ($\mathrm{TP}^c_{B_{ps}}$) and false positives (FP^c). The total number of true positives is defined as ($\mathrm{TP}^c = \mathrm{TP}^c_{B_{gt}} + \mathrm{TP}^c_{B_{ps}}$). 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:

$$\mathrm{TP}^c_{B_{gt}} = \sum_{x \in B^c_{gt}} z \text{ with } z = \begin{cases} 1 - (d(x, S^c_{ps})/\theta)^2 & \text{if } d(x, S^c_{ps}) < \theta \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

$$\mathrm{FN}^c = |B^c_{gt}| - \mathrm{TP}^c_{B_{gt}} \tag{13}$$

$$\mathrm{TP}^c_{B_{ps}} = \sum_{x \in B^c_{ps}} z \text{ with } z = \begin{cases} 1 - (d(x, S^c_{gt})/\theta)^2 & \text{if } d(x, S^c_{gt}) < \theta \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

$$\mathrm{FP}^c = |B^c_{ps}| - \mathrm{TP}^c_{B_{ps}} \tag{15}$$

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

$$\mathrm{BJ}^c = \frac{\mathrm{TP}^c}{\mathrm{TP}^c + \mathrm{FP}^c + \mathrm{FN}^c}. \tag{16}$$

This new score is not zero when the ground truth and the predicted segmentation for a given class have some

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	← **JI** →	.88	.77	.66	.54	.30
0	0	0	0	.99	← **BF** →	.99	0	0	0	0
0	.20	.46	.47	.77	← **BJ** →	.79	.64	.50	.50	.48

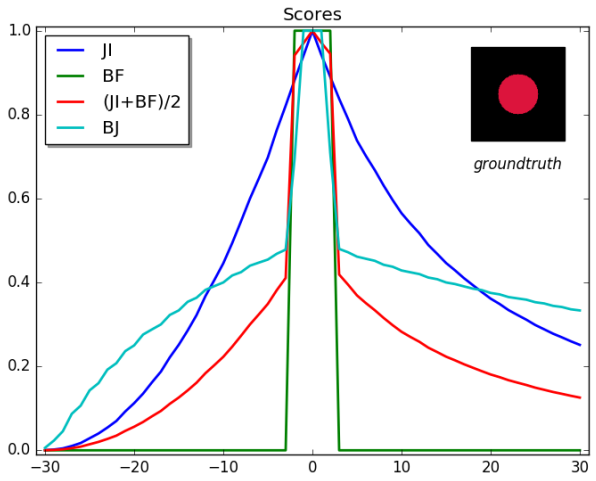


Fig. 2: Per-class scores of the segmented circle (top-right) for different levels of infra/over segmentation. The parameter θ is set to 4 pixels for both BF and BJ, which corresponds to 0.0075 of the image diagonal.

overlapping ($|S^c_{gt} \cup S^c_{gt}| > 0 \Rightarrow \text{BJ}^c > 0$). This behavior is similar for the metric JI^c but not for BF^c. On the other hand, the BJ^c score 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. Figure 2 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 showed 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. The score for a given image sequence is obtained as the average of per-image BJ's over the number of images contained in the sequence. It is worth to mention that per-image scores are more interesting than scores obtained over the full dataset (i.e., where a single BJ^c score is computed) for several reasons, as discussed in [19]. To mention some of these: *i)* per-image scores reduce the bias *wrt.* very large objects, and *ii)* they allow the statistical analysis about the performance of different segmentation frameworks in different parts of the dataset.

IV. EXPERIMENTAL ANALYSIS OF ACCURACY METRICS

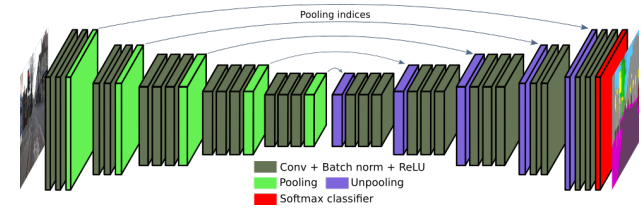
This section presents a number of qualitative and quantitative results showing the accuracy of different types of CNN trained and tested on the Virtual KITTI [12] and KITTI [17] datasets and the comparison of the different evaluation metrics. The results confirm that measuring accuracy in the neighborhood 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 the performance of these two scores. Note that in the following experiments, we focus our attention to the point of evaluating different accuracy metrics as it's the aim of this paper, rather than evaluating the suitability of different network architectures to the problem of semantic segmentation in urban scenes.

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

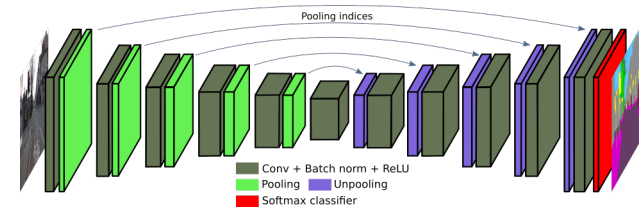
Using color and depth information has proven to be useful for semantic segmentation [2], [14], [8]. However, it's not clear yet how these two types of data should be fed into the CNN, and which network architecture is optimal for the problem. Without trying to solve this problem, we just describe here several solutions in order to compare later the suitability of different accuracy statistics. The network models analyzed in the next section are FuseNet [8], SegNet [6], and some modified versions of the latter that we describe here.

We introduce a modification of the VGG16 topology [24] employed by SegNet (see fig. 3a) to obtain a more compact network which we call Compact Encoder Decoder Convolutional Neural Network (CEDCNN), which is illustrated in fig. 3b. This network model increases the number of parameters of the filters in each resolution to produce higher dimensional feature maps, and reduces the number of consecutive convolution filters (convolution+batch normalization+ReLU) to reduce the complexity and non-linearity of the model. 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 as shown in figure 3c. In the same way as for SegNet2, we also modify our model CEDCNN to obtain a new network, called

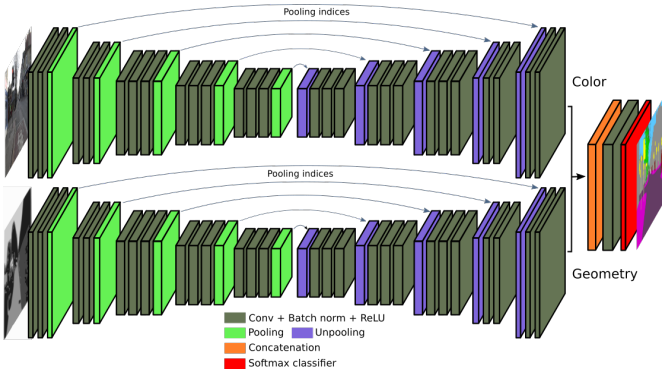
CEDCNN2, with two different pipelines to extract feature maps from color and geometric information separately.



(a) SegNet (VGG16 without Fully SegNetnnected layers)



(b) CEDCNN (60% less parameters than SegNet)



(c) SegNet2 (Color+Geometry)

Fig. 3: CNN topologies employed in our experiments: a) SegNet, b) CEDCNN, c) SegNet2.

B. Comparison of different metrics

Firstly, we provide a qualitative analysis of the behavior of different metrics with infra-segmented and over-segmented objects, as shown in fig. 2. We produce synthetic segmentations of the ground truth of different object classes of interest, 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 labeled pixels of 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. Figure 2 shows the score of different per-class metrics: JI, BF, the average of JI and BF, and BJ. 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 see that the Jaccard index has the most gradual behavior, since it depends only on the amount of pixels correctly and wrongly classified. The BF score measures the quality of the segmented boundaries, it shows a discontinuous trend according to the threshold parameter used to distinguish inliers from outliers. The previous measures may be 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 metric is less severe than for BF, it is still something undesirable because the score depends highly on the threshold value θ. Finally, the BJ score shows a continuous behavior because it's value depends on the distance, instead of a filter, so that the θ parameter has less influence on its value. The BJ score is higher than JI for infra-segmented objects, which is interesting because to avoid miss-classifications. The BJ score is close to 0.5 for over-segmented objects with bad contour segmentation. This effect 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.

C. Semantic segmentation of RGB and Depth on Virtual KITTI

This experiment makes use of the Virtual KITTI dataset [12] for training and testing different models for semantic segmentation. This 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 [17], 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 the same dataset.

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 [2].

Table III presents the accuracy measured as the recall (R), the mean recall of all classes, the mean JI and considering the metric BJ. The best scores are highlighted in bold. The first 5 rows of the table (white background) 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 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) [1]. Regarding the accuracy of the model SegNet2 (see fig. 3c), the use of input data from RGB-ND achieves the best results, for which all the global accuracy metrics indicate that it is the best model. Note that recall measured on the class borders are very close to the mean recall. In fact, both measures are quite similar because computing the recall only on class borders leverages the effect of unbalanced frequencies of the different classes, while being more stable to the presence of low-frequency ("rare") classes with lower class-wise accuracy.

TABLE III: Semantic segmentation accuracy of SegNet and SegNet2 using color and geometric information (in %).

Model \ Metric	recall	m. R	m. JI	BJ
SegNet (RGB)	81.7	61.9	41.2	61.7
SegNet (D)	85.8	65.2	47.0	67.1
SegNet (ND)	**88.6**	70.2	**51.1**	**69.8**
SegNet (NE)	88.5	**71.5**	48.9	69.5
SegNet (RGBD)	78.1	64.1	41.8	60.7
SegNet2 (RGB-D)	88.5	71.0	49.4	70.5
SegNet2 (RGB-ND)	**90.3**	**71.8**	**52.9**	**71.7**

We analyze next other network architectures like FuseNet [8], together with the network topologies introduced in section IV-A: SegNet2, CEDCNN and CEDCNN2. Table IV shows the accuracy measured with the same global statistics of the previous table. For easier reference, this table also shows the results of SegNet for RGB and SegNet2 for RGB-ND in the two first rows. 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 argued in [8] 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. For this case, the images are "virtually" acquired from a forward facing camera mounted in a car. Therefore, the surface directions have some invariants, such as the angle with gravity, that constitute a relevant source of information.

TABLE IV: Global accuracy of different types of networks using color and geometric information (in %).

Model \ Metric	recall	m. R	m. JI	BJ
SegNet (RGB)	81.7	61.9	41.2	61.7
SegNet2 (RGB-ND)	88.6	70.2	51.1	69.8
FuseNet (RGB-D)	85.2	65.9	45.8	64.9
FuseNet (RGB-ND)	88.1	64.6	47.2	68.9
CEDCNN (RGB)	88.8	72.8	48.6	70.5
CEDCNN2 (RGB-D)	90.1	79.7	60.0	77.5
CEDCNN2 (RGB-ND)	**92.6**	**81.7**	**64.7**	**80.0**

We remark that the different models achieve the best semantic segmentation depending on the class, while the best model overall (according to BJ) is CEDCNN2 with RGB-ND, which has a considerable better performance segmenting classes with lower frequencies, such as "traffic light" or "truck", while the scores of large frequency classes like "sky", "tree" or "road" are generally more stable across the different models. This fact is depicted in fig. 4 with confusion matrices for three different architectures. Note that if we need to choose between one of the FuseNet models, we need to consider the metric for all classes. Having unbalanced class frequencies has a great influence on the final score, because multi-resolution CNN are well suited by design to segment large homogeneous classes, but they are harder to train in order to achieve similar scores on low frequency classes, which sometimes are more important for many practical applications like for the case of autonomous driving.

Regarding the different accuracy metrics, we observe that the mean recall and the mean JI are less stable across the different experiments. This occurs because the accuracy of low frequency classes have a large variability even for similar models, and this variability is also reflected in their mean values. This effect is also observed in the normalized confusion matrices, see fig. 4, where the diagonal elements correspond to recall of each class, and where the JI for the i-th class is related to the values contained the i-th row and i-th column. 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.

V. CORRELATION OF DIFFERENT METRICS

This section measures the correlation of the different metrics evaluated in the previous experiment. 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 Spearmans rank correlation (ρ), which is a nonparametric measure of rank correlation, defined as the Pearson correlation coefficient between the ranked variables. It is used here to measure 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

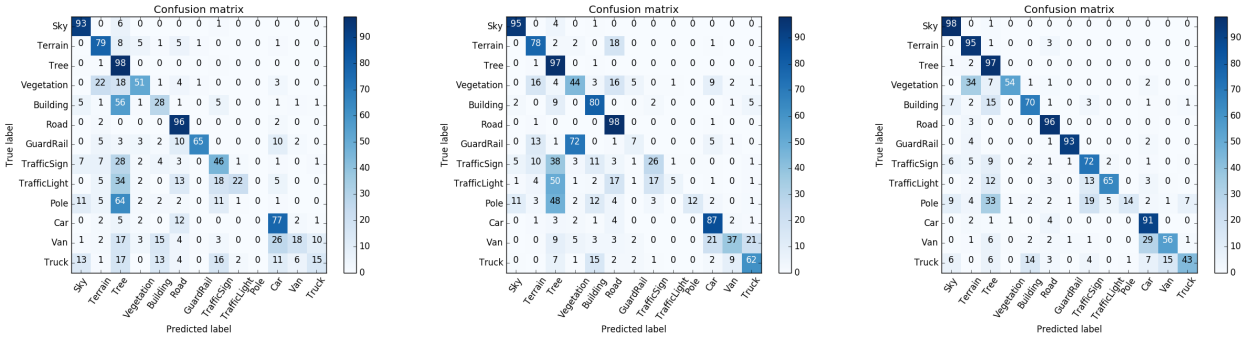
$$\rho = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}} \tag{17}$$

where rg_X, rg_Y are the ranks of the score distributions X, Y. Since we choose integer values for the rank, the formula is simplified to

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(rg_{X_i} - rg_{Y_i})^2}{n(n^2-1)} \tag{18}$$

Table V shows the ranking correlations among metrics, where we can see that the BJ score is correlated to both JI and BF, showing that they capture similar information. Notice that the correlations with BJ are higher than the correlations among other pairs of scores.

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



(a) SegNet (RGB) (b) FuseNet (RGB-D) (c) CEDCNN2 (RGB-ND)

Fig. 4: Normalized confusion matrices (in %) of semantic segmentation in the real KITTI dataset with: a) SegNet (RGB), b) FuseNet (RGB-D) and c) CEDCNN2 (RGB-ND).

TABLE V: Spearmans rank correlation of different segmentation scores.

metric	JI	BF	(JI+BF)/2	BJ
JI	-	0.48	0.59	**0.63**
BF	-	-	0.68	**0.65**
(JI+BF)/2	-	-	-	0.73

VI. 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. The global recall, mean recall and mean JI statistics have been traditionally employed to evaluate different image segmentation results, however, these metrics are not satisfactory enough when the classes frequencies are very unbalanced. We present a simple and efficient strategy to compute the recall on border regions of the different classes which leverages unbalanced frequencies, and is a good indicator to measure class segmentation. Our proposed metric encodes jointly the rate of correctly labeled pixels and how homeomorphic is the segmentation to the real object boundaries. We also present results for several different CNN architectures using two state-of-the-art benchmark datasets. Though we address this problem in the context of urban images segmentation, our results can also be extended to other contexts, like for indoor scenarios.

The research in this paper was partly motivated by the need of segmentation solutions with better segmentation of contours, for which traditional metrics were not suitable. In our future research, we plan to study how to give more importance to the segmentation of such contours during the training phase of the CNN and on obtaining optimal CNN designs for semantic segmentation of complex dynamic outdoor scenes.

REFERENCES

[1] R. Drouilly, P. Rives, and B. Morisset, "Semantic representation for navigation in large-scale environments," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1106–1111.

[2] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision*. Springer, 2014, pp. 345–360.

[3] 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.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[5] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[7] 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.

[8] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Proc. ACCV*, vol. 2, 2016.

[9] L. Ladickỳ, P. Sturgess, 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.

[10] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[11] 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.

[12] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4340–4349.

[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 *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 681–687.

[15] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.

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

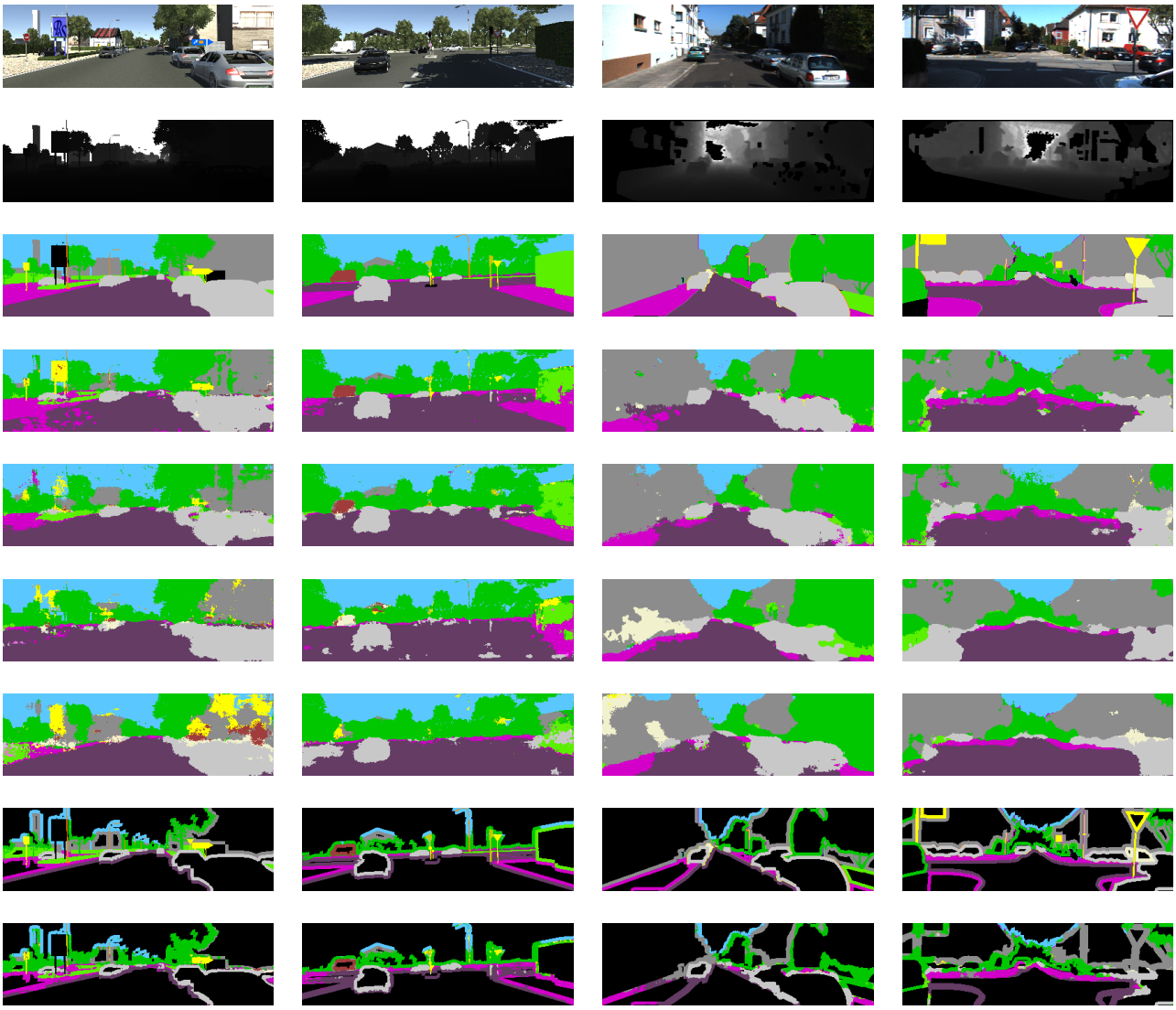


Fig. 5: Semantic segmentation produced by the different models. The first 2 columns correspond to test images from Virtual KITTI, while and the last 2 columns correspond to images from our KITTI test. The first row shows the input RGB image, followed by depth, groundtruth labels. Rows 4^{th} to 7^{th} show the segmentation produced by: CEDCNN2 (RGB-D), CEDCNN (RGB), SegNet2 (RGB-D) and SegNet (RGB) respectively. The 8^{th} row shows the border regions from the ground truth, which are used to evaluate border recall and our metric in eq. (16). The 9^{th} row shows the border precision of CEDCNN2 (RGB-D).

[17] 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.

[18] 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.

[19] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?." in *BMVC*, vol. 27, 2013, p. 2013.

[20] P. Kohli, L. Ladicky, and P. H. Torr, "Robust higher order potentials for enforcing label consistency," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[21] J. Freixenet, X. Muñoz, D. Raba, J. Martí, and X. Cufí, "Yet another survey on image segmentation: Region and boundary information integration," *Computer VisionECCV 2002*, pp. 21–25, 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.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.