Weighted Intersection over Union (wIoU): A New Evaluation Metric for Image Segmentation

Yeong-Jun Cho
Department of Artificial Intelligence Convergence, Chonnam National Universiy
yj.cho@chonnam.ac.kr

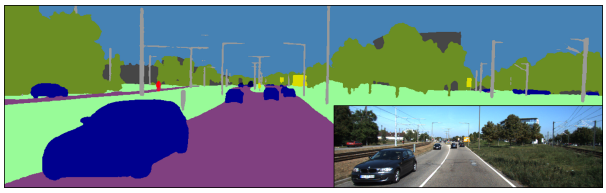
Abstract

In recent years, many semantic segmentation methods have been proposed to predict label of pixels in the scene. In general, we measure area prediction errors or boundary prediction errors for comparing methods. However, there is no intuitive evaluation metric that evaluates both aspects. In this work, we propose a new evaluation measure called weighted Intersection over Union (wIoU) for semantic segmentation. First, it builds a weight map generated from a boundary distance map, allowing weighted evaluation for each pixel based on a boundary importance factor. The proposed wIoU can evaluate both contour and region by setting a boundary importance factor. We validated the effectiveness of wIoU on a dataset of 33 scenes and demonstrated its flexibility. Using the proposed metric, we expect more flexible and intuitive evaluation in semantic segmentation field are possible.

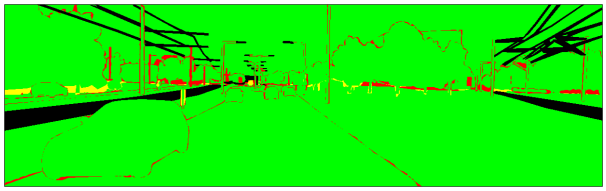
1. Introduction

In recent years, many studies have tried to train pixel-level classifiers on large-scale image datasets to perform accurate semantic image segmentation [27, 31]. The goal of the image segmentation is to assign a class label of each pixel in the scene. It is one of the most basic problems in computer vision fields and has various potential applications e.g., autonomous driving systems [8], scene understanding [21] and medical image diagnostics [15, 17]. Recently, significant advances in deep learning techniques are driving high-performance of semantic segmentation in various benchmark datasets [6, 9, 13]. Figure 1 shows a result of the semantic segmentation [31] and its error map in the KITTI dataset [13].

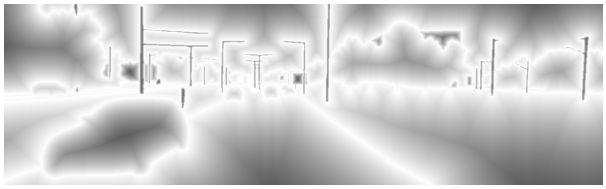
Although the method performs the best compared to other state-of-the-art methods, it still experiences classification errors. Most of the errors (i.e., false negatives and positives) occur around object boundaries due to high classification ambiguities around the boundaries as shown in Fig. 1 (b). That is because at least two classes are adjacent around the objects' boundaries making class predictions more difficult and ambiguous. Even though the proportion of boundaries in the scene is very small, it has a significant impact on qualitative results. In contrast, the internal areas of objects have relatively low classification ambiguities and occupy most of the scene area.



(a) Segmentation result [31] with an original image



(b) Error map. Color of each pixel denotes: (green) True positive, (red) False negative, (yellow) False positive, (black) No label.



(c) Proposed weight map for the segmenation evaluation

Figure 1. Challenges in semantic segmentation and the proposed weight map. (a–b) It is difficult to predict pixel labels around object boundaries. (c) The proposed weight map emphasizes the importance of boundaries.

In recent years, many studies have focused on the importance of the boundary and tried to enhance the classification accuracy based on the boundary information [3, 22, 27].

They added additional network branches in baseline networks and optimized boundary loss terms to leverage the boundary cues. To evaluate semantic segmentation results, most methods [27, 31] rank their methods by measuring intersection-over-union (*IoU*) score described in Section. 3.2. Unfortunately, *IoU* evaluates all pixels equally. Performance scores improve very subtly according to the evaluation criteria. To directly evaluate the boundary estimation results, several evaluation metrics such as BF score [7] and Boundary Jaccard (BJ) [12] were proposed; however, they do not evaluate internal areas of the objects. In addition, those methods are highly dependent on the setting of distance tolerance parameter θ. As we mentioned above, both evaluations of internal and boundaries of objects are important, without relying on certain tuning parameters.

In this work, we propose a novel evaluation measure for semantic segmentation called weighted Intersection over Union (*wIoU*). It generates a weight map based on the boundary distance map, as shown in Figure. 1 (c). The weight map is adjustable based on a boundary importance factor α, as shown in Figure. 5. Using the weight map enables weighted evaluation for each pixel. Setting the value of α is intuitive. Furthermore, it effectively evaluates both boundary and internal regions simultaneously. Our work contributes in the following ways.

- Can evaluate both contour and region with a single evaluation metric
- Does not depend on many hyper-parameters
- Validated the proposed metric in the various test images

For comparing evaluation metrics, we construct a simple semantic segmentation dataset comprising 33 scenes. We have validated that the proposed *wIoU* can evaluate both contours and regions according to the boundary importance factor α. When the α value is small, it functions similarly to region-based evaluation metrics, particularly *IoU*; otherwise it functions similar to boundary-based evaluation metrics. We anticipate that the proposed *wIoU* will be used not only for evaluating metric but also as a novel region-contour loss for training accurate semantic segmentation networks.

2. Problem definition

Image segmentation (also called semantic image segmentation) is a problem to find several sub-regions or partitions within the given image in the fields of computer vision and digital image processing. The sub-regions consist of sets of pixels belonging to objects. The goal of image segmentation is to predict the class labels $\mathbf{C}$ of pixels in the given image $\mathbf{I}$. An image can be represented as a set of pixels as follows,

$$\mathbf{I} = \{i_p | p = (p_x, p_y), 1 \leq p_x \leq \mathrm{W}, 1 \leq p_y \leq \mathrm{H}\}, \quad (1)$$

where i_p denotes the value of pixel p. W and H represent the width and height of the image, respectively. In general, an image consists of 3-channel (e.g., R,G,B), but for convenience, we do not consider the channel index.

Figure 2. Example of an image and its predicted classes

Using the given image pixels i_p, the probability distribution of class for each pixel is predicted as $p(c|i_p)$. To predict the probability distribution, we can employ various semantic segmentation approaches [27, 31]. Next, we select the class label for each pixel based on the distributions as follows,

$$c_p^* = \underset{c \in \mathcal{C}}{\mathrm{argmax}} p(c|i_p), \quad (2)$$

where $\mathcal{C}$ is a set of class (e.g., car, tree, road and Etc). Finally, we build a predicted class map of the given image as

$$\mathbf{C} = \left\{c_p^* | p = (p_x, p_y), 1 \leq p_x \leq \mathrm{W}, 1 \leq p_y \leq \mathrm{H}\right\}. \quad (3)$$

Then, $\mathbf{C}(p)$ denotes the class label c at the point $p = (p_x, p_y)$. Figure 2 shows an example of a semantic segmentation result with predicted classes. As a result of semantic segmentation, each pixel corresponds to a distinct class label.

3. Related works

3.1. Semantic Image Segmentation Methods

Semantic image segmentation is one of the most important problems in the field of computer vision research. With the recent development of deep learning, many attempts to apply deep learning to semantic segmentation applications have been presented. R.Girshick *et al*. [14] proposed region proposal convolutional neural networks called R-CNN for object detection and semantic image segmentation. In addition, J. Long *et al*. [20] also proposed a semantic segmentation method via fully convolutional networks (FCN) which is an end-to-end framework performing pixelwise prediction. Thanks to the large model capacity of deep neural networks, they significantly improved the prediction

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

Table 1. Confusion matrix for classification. Blue color indicates correct estimation and red color indicates failure estimation.

performance compared to many popular traditional methods [4, 11, 18, 25, 26] reviewed in [28].

Recently, many methods [3,5,19,22,27,29] have focused on the importance of boundary information in the scene to achieve accurate semantic segmentation. They added additional network branches and optimized the boundary loss to capture the boundary information. Then, the estimated boundary cues in the scene are effectively used for semantic segmentation. Many authors have argued that the boundary cue alleviates the problem of boundary blur and leads to edge-preserving results. Various benchmark datasets such as PASCAL VOC [9], Cityscapes [6], ADE20K [30], ISPRS2D [1] and KITTI [2, 13] are commonly used to train and evaluate the methods.

3.2. Evaluation Metrics

• **Basic notations for classification results.** To evaluate the results of classification, there are four basic notations according to the actual class and predicted class: (1) True Positive (TP); (2) True Negative (TN); (3) False Positive (FP); (4) False Negative (FN). Each notation is summarized in Table. 1. TP and TN are correct estimations. FP and FN are false alarms, which are called Type I error and Type II error, respectively.

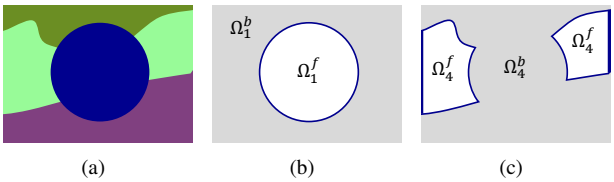
• **Region-based evaluation.** Based on the types of classification results, many studies in fields such as machine learning and pattern recognition have evaluated their methods. For example, the evaluation metrics such as Precision and Recall are very widely used: $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, Furthermore, F_1 score, which is the weighted average of *Precision* and *Recall*, has been used as an evaluation metric:

$$F_1 = \frac{2TP}{2TP + FN + FP}, \tag{4}$$

It reflects two evaluation results (*Precision* and *Recall*) evenly. Another common metric for evaluating classification is a Jaccard index (*JI*):

$$JI = \frac{TP}{TP + FN + FP}, \tag{5}$$

For semantic image segmentation problem defined in Section. 2, Intersection over Union (*IoU*) is equivalent to *JI*:



(a) (b) (c)

Figure 3. (a) Example of a ground-truth class map $\mathbf{C}_{gt}$. (b,c) Corresponding foreground (Ω_c^f, white area) and background (Ω_c^b, shaded area) regions where $c = 1$ and $c = 4$, respectively. Solid blue line is the boundary of the object.

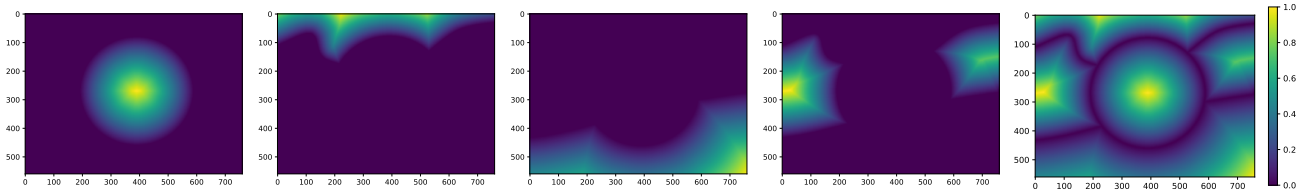
$$IoU = JI = \frac{|\mathbf{C} \cap \mathbf{C}_{gt}|}{|\mathbf{C} \cup \mathbf{C}_{gt}|}, \tag{6}$$

where $\mathbf{C}$ denotes a predicted class map and $\mathbf{C}_{gt}$ denotes a ground-truth class label map. Most of the previous studies related to semantic segmentation [27, 31] have adopted *IoU* measure. It compares a similarity between two regions ($\mathbf{C}$ and $\mathbf{C}_{gt}$) to evaluate all predicted pixels of the image.

• **Boundary-based evaluation.** Meanwhile, many studies [7, 24] have pointed out that the prediction performance on the boundaries of objects has a significant impact on the perceived segmentation quality. In fact, the object boundary has higher prediction difficulty than the internal area of the object – at least two classes are adjacent around the objects' boundaries, which makes class prediction more difficult and ambiguous. However, *IoU* equally evaluates all pixels in the image, making it hard to evaluate performances around object boundaries.

Some studies [23, 24] evaluated the edge prediction results by measuring edge-based F_1 score as in Eq. 4 only around the objects' edges via bipartite matching. To make the evaluation metric robust to small errors, they set the distance tolerance (θ) for performing bipartite matching between predicted edge points and ground-truth edge points. In addition, BF score [7] and Boundary Jaccard (BJ) [12] were also proposed to capture the edge prediction results.

The evaluation metrics can measure the boundary prediction performance; however they still require certain distance tolerance values (θ), which determine the boundary region in the scene. Selecting a distance tolerance (θ) leads an inflexibility in evaluation measures. Furthermore, region-based evaluation metrics and boundary-based evaluated metrics are separate, making it difficult to compare the image segmentation performances at a glance. In this work, we propose a single evaluation metric that can measure both evaluation factors (i.e., region- and boundary-based) in Section. 4.



(a) $\bar{\mathbf{D}}_1(p)$ (b) $\bar{\mathbf{D}}_2(p)$ (c) $\bar{\mathbf{D}}_3(p)$ (d) $\bar{\mathbf{D}}_4(p)$ (e) $\bar{\mathbf{D}}(p)$

Figure 4. Examples of distance maps. The ground-truth class map is illustrated in Figure 3 (a).

4. Proposed Evaluation Metric

4.1. Motivation and main ideas

We observed that it is more difficult to infer the label of the boundary of each object. This is because the inference probability of each label is mixed at the boundary. Unfortunately, the details of an object are usually at the boundary. Even if most of the object region is predicted correctly, if it fails to predict the boundary of the object, it looks qualitatively bad. In addition, missing object details can lead to poor results in applications of semantic segmentation.

To evaluate segmentation results while considering these issues, we first generate a weighted map based on a distance transform in Section. 4.2. The weight map emphasizes the importance of the object's boundaries. We then calculate the weighted intersection over union (*wIoU*) using the generated weight map in Section. 4.3.

4.2. Weighted map generation

The problem of a distance transform is to compute the distance of each point from the boundary of foreground region [10]. First, it assigns fore- and back-ground regions according to the class label c as

$$\Omega_c^f = \{p|\mathbf{C}_{gt}(p) = c\}, \Omega_c^b = \{q|\mathbf{C}_{gt}(q) \neq c\}. \quad (7)$$

Each region consists of a set of points. Figure 3 shows predicted map ($\mathbf{C}$) (Figure. 3 (a)) and its corresponding fore- and back-ground regions (Ω_c^f, Ω_c^b) where $c = 1$ and $c = 4$ (Figure. 3 (b) and (c)). Solid blue lines are the boundaries of objects.

Next, it generates a distance map that consists of a distance value of each point p by computing the smallest distance from the back-ground region Ω_c^b. It is represented by

$$\mathbf{D}_c(p) := \min\left\{\mathrm{d}(p,q)\,|p \in \Omega_c^f, q \in \Omega_c^b\right\}, \quad (8)$$

where $\mathrm{d}(p,q)$ is the pixel distance between two points p and q. To compute the distance, we can adopted ρ-norm (also called l_ρ norm) defined by

$$\begin{aligned}\mathrm{d}(p,q) :=& \|p-q\|_\rho \\ =& \left(|p_x - q_x|^\rho + |p_y - q_y|^\rho\right)^{\frac{1}{\rho}},\end{aligned} \quad (9)$$

where $\rho \geq 1$. ρ-norm generates various distance metrics such as Manhattan ($\rho = 1$), Euclidean ($\rho = 2$), and Chessboard ($\rho = \infty$) distances according to its value ρ. In this work, we set ρ as 2 to follow Euclidean distance for generating the weight map $\mathbf{D}_c$. As it mentioned in Section 4.1, we aim to generate a weight map according to the distance from the object boundary. Calculating Euclidean distances can efficiently approximate to compute the normal distances from the boundaries. As a result, the distance maps for all categories $\mathcal{C}$ can be simply calculated.

However, the distance map of each object has different distance range due to the difference in size between the objects. It causes unfair performance evaluation in many cases. To handle this, we designed the regularized distance as

$$\bar{\mathbf{D}}_c(p) = \frac{\mathbf{D}_c(p)}{\max\left(\mathbf{D}_c(p)\right)}, \quad (10)$$

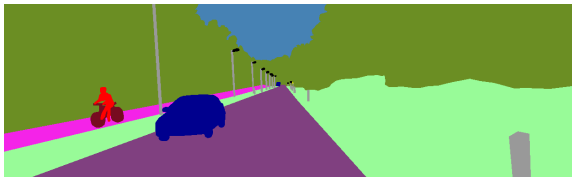
Then it lies on $[0,1]$ for all pixel points p and classes c. Note that if a class c has multiple instances, the regularization in Eq. 10 is performed for each instance. Figure. 4 shows examples of regularized distances. Considering all the class in the class map $\mathbf{C}$, the distance maps $\bar{\mathbf{D}}_c$ can be combined into one distance map as follows

$$\bar{\mathbf{D}}(p) = \left\{\bar{\mathbf{D}}_c(p)\,|1 \leq c \leq \mathcal{C}\right\}, \quad (11)$$

An example of the combined distance map is shown in Figure. 4 (e). Finally, a weight map is calculated by

$$\mathbf{W}(p) = e^{-\alpha\bar{\mathbf{D}}(p)}, \quad (12)$$

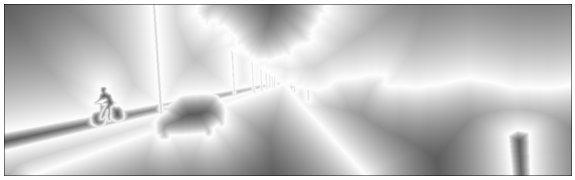
where $\alpha > 0$ is a boundary importance factor. As shown in Figure. 5, the value of α determines the importance of boundaries in the scene. When the boundary importance is small ($\alpha = 0.01$), the weight map is uniformly distributed (Figure. 5 (b)). It means that all areas have about the same weights. On the other hand, when the boundary importance is comparatively large ($\alpha = 100$), it emphasizes boundaries in the scene (Figure. 5 (f)). In this way, we can set the weight map very simply by manipulating only one parameter α.



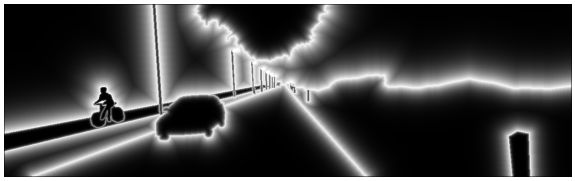
(a) Ground truth $\mathbf{C}_{gt}$

(b) Weight map $\mathbf{W}$, $\alpha = 0.01$

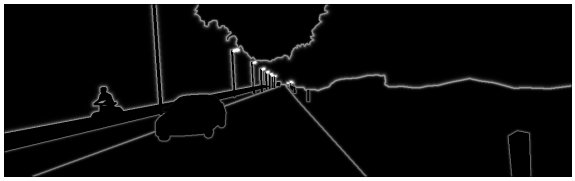
(c) Weight map $\mathbf{W}$, $\alpha = 0.1$



(d) Weight map $\mathbf{W}$, $\alpha = 1$



(e) Weight map $\mathbf{W}$, $\alpha = 10$



(f) Weight map $\mathbf{W}$, $\alpha = 100$

Figure 5. Ground truth and its weight maps according to the boundary importance factor α. The range of weight map is $[0, 1]$. White and black pixels denote weight values $1, 0$, respectively.

(a) Scene #01 (b) Scene #02 (c) Scene #03

Figure 6. Examples of test scenes and their segments.

4.3. Weighted intersection over union

Based on the calculated weight map in Section. 4.2, we propose a new evaluation metric called Weighted Intersection over Union (*wIoU*) by

$$wIoU = \frac{|\mathbf{C} \cap (\mathbf{C}_{gt} \circ \mathbf{W})|}{|\mathbf{C} \cup (\mathbf{C}_{gt} \circ \mathbf{W})|}, \tag{13}$$

where $\circ$ is an element-wise product called Hadamard product [16]. In this step, we omit p for convenience and all terms in the Eq. 13 have the same dimension (i.e., size of an input image). Comparing the proposed *wIoU* with conventional *IoU* (Eq. 6), it emphasizes the importance of each pixel based on the weight map $\mathbf{W}$.

5. Experimental Results

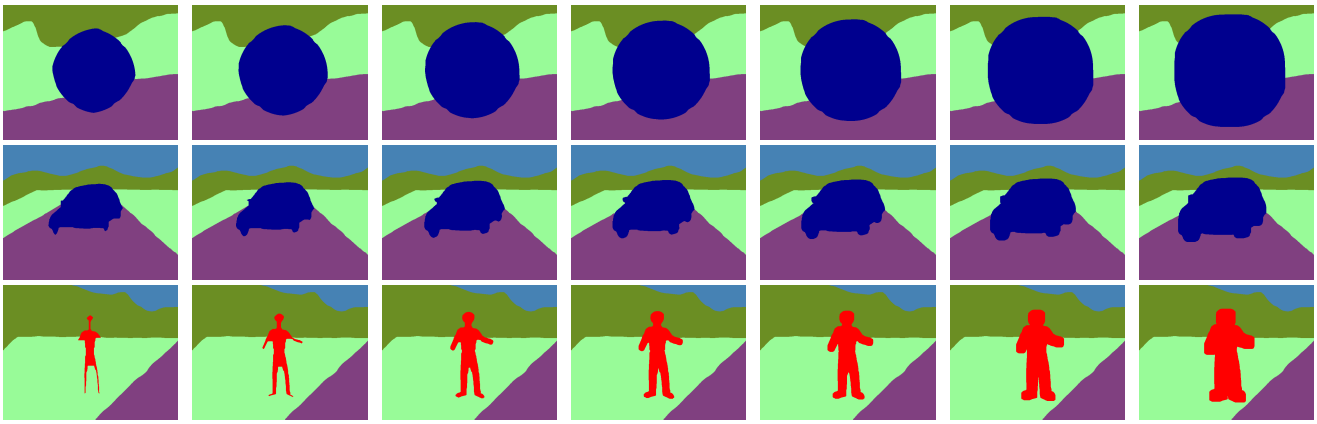
5.1. Datasets

To compare evaluation metrics, we built a segment dataset as shown in Figure. 6. We followed the `KITTI` [13] color assignments for classes. At the center of each scene, we placed an object (e.g., car, person). For each scene, we created test segments (Figure. 7 (d)) that include prediction errors around boundaries, which commonly occur. To consider more possible situations, we applied morphological image processing to the objects in test images. We performed 5 different levels of erosion and dilation for each image. As a result, the total number of segment images is 33. These test images will be used to compare overall similarities between different evaluation metrics in Section. 5.2.

Additionally, we created three test segment images that contain the same number of error pixels (FP+FN), but the positions of the error pixels are different, as illustrated in Figure. 9. These test scenarios are quite unusual in semantic segmentation. However, we can precisely compare how each evaluation metric works under the same number of error pixels.

5.2. Comparisons of different metrics

In this section, we compare three different evaluation metrics: 1) Intersection over union (*IoU*), 2) Proposed weighted IoU (*wIoU*), 3) edge-based score F_1 [23]. For the proposed *wIoU*, we tested five different levels of boundary



(a) Erode lv5 (b) Erode lv3 (c) Erode lv1 (d) Test segments (e) Dilate lv1 (f) Dilate lv3 (g) Dilate lv5

Figure 7. Examples of generated test segments. (d) Generated test segments for each scene including prediction error. Based on morphological image processing (5 different levels of erosion (a–c) and dilation (e–g)), the test segments were augmented.

importance factors α as 0.01, 0.1, 1, 10 and 100. For the edge-based score F_1, we set the distance tolerance value θ as 3 pixels. To compare the evaluation results of the metrics, correlations $\rho_{X,Y}$ and sum of difference $D_{X,Y}$ are calculated as follows:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, \tag{14}$$

$$D_{X,Y} = E[|X - Y|], \tag{15}$$

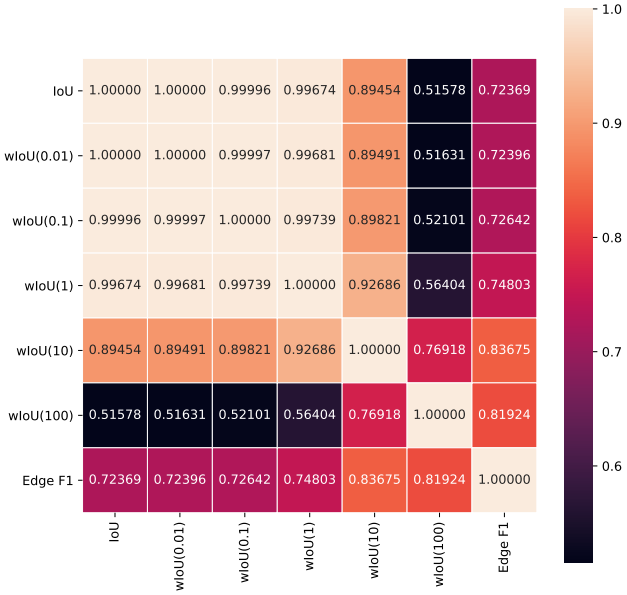
where X, Y are sets of evaluation values calculated by each evaluation metric, $cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$ is a covariance matrix between X and Y. μ and σ are mean and standard deviations of the evaluation values, respectively.

The correlation $\rho_{X,Y}$ measures strength of a relationship between two variable sets X, Y. Thus, although the absolute values between two variable sets are different, they have a correlation value when the tendency of value change is similar. The value of correlation lies on $[0, 1]$, and the value 1 indicates the high similarity. On the other hand, sum of difference $D_{X,Y}$ measures the absolute value difference between the two variable sets. It can lie on $[0, \infty]$, and the value 0 indicate the high similarity. Based on both evaluations, we can measure a relationship and differences between metrics.

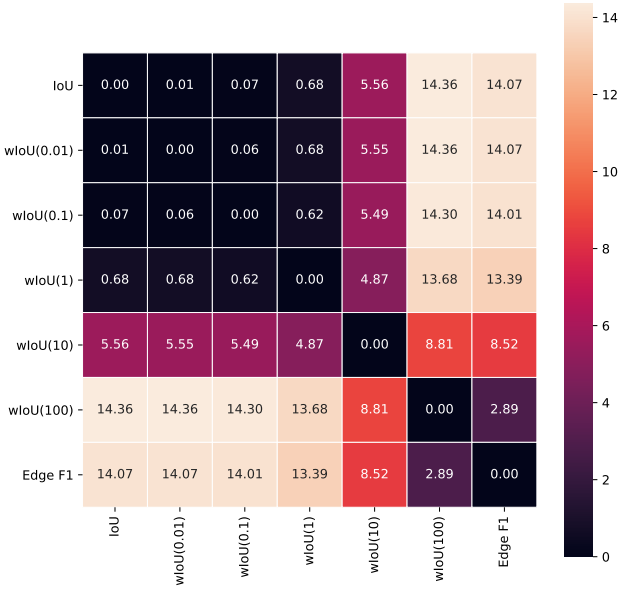
Figure. 8 illustrates correlations and summations of differences between *IoU*, *wIoU*, and edge-based F_1. *wIoU* with low boundary importance factor ($\theta = 0.01$) is very similar with *IoU* in terms of both correlation and sum of differences. On the other hand, *wIoU* with high boundary importance factor ($\theta = 100$) shows the high similarity with edge-based F_1. The correlations of edge-based F_1 between other metrics are not meaningful differences, while the sum of differences between edge-based F_1 and *wIoU* series become lower as the boundary importance factor increases. In this way, the proposed *wIoU* combines both conventional region-based and boundary-based evaluation metrics depending on the boundary importance factor θ.

For further comparisons of evaluation metrics, we additionally tested three test images as illustrated in Figure. 9. Those test images were generated from Scene #01 (Figure. 6 (a)) while they keeps the same numbers of error pixels (FP+FN). Therefore region-based *IoU* metric showed the same evaluation results (0.972) for all test images. On the other hand, other evaluation metrics (i.e., *wIou*, edge-based F_1, edge-based recall) measure different aspects of performance evaluations. *wIoU* with small boundary importance factors ($\theta = 0.01, 0.1$) are very relevant to *IoU* metric as they worked in common test images in Figure 8. When boundary importance factor was set large ($\theta = 100$) for *wIoU*, it works almost the same with edge-based Recall ($\frac{TP}{TP+FN}$). This result indicates that *wIoU* with large boundary importance is difficult to measure false positive (FP) pixels which are not around ground-truth boundaries. An edge-based F_1 was highly dependent on the predicted boundary in the scene; it evaluated very different values for each test image ($0.5843, 0.936, 0.8603$) despite of the same error pixels.

wIoU with a boundary importance factor ($\theta = 1$) makes difference between the test images. It gives more weight to the boundaries, but it can evaluate both the regions and the boundaries in a reasonable way. Based on the experimental results, we expect that the proposed *wIoU* will not only be utilized as a reasonable evaluation metric but also as a new region-contour loss for training accurate semantic segmentation networks.



(a) Correlations

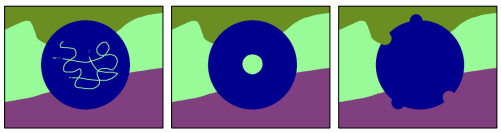


(b) Summations of differences

Figure 8. Comparison matrices of different metrics

6. Conclusions and Future Works

In this work, we proposed a new evaluation measure (*wIoU*) for semantic segmentation. First, we proposed a weight map generation from a boundary distance map, allowing weighted evaluation for each pixel based on a boundary importance factor. The proposed *wIoU* can evaluate both boundary and region by setting a boundary importance factor. We validated the effectiveness of *wIoU* on a dataset of 33 test images and demonstrated its flexibility and rationality.



IoU	0.9720	0.9720	0.9720
wIoU(0.01)	0.9721	0.9722	0.9720
wIoU(0.1)	0.9727	0.9734	0.9713
wIoU(1)	0.9779	0.9837	0.9649
wIoU(10)	0.9981	0.9999	0.9159
wIoU(100)	1.0000	1.0000	0.8808
Edge F_1	0.5843	0.9360	0.8603
Edge Recall	1.0000	1.0000	0.8833

Figure 9. Evaluations of different metrics on specific test images

Based on our study, there are a few areas that could be explored for the future works. Beyond using it as an evaluation metric, we can adopt the proposed *wIoU* as an optimization loss for deep neural networks. To demonstrate the efficiency of *wIoU* in training networks, we have to design various structures of deep neural networks and experimentally demonstrate performance improvements in the next works.

References

[1] Isprs 2d semantic labeling contest. In *Available online: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html*, 2018. 3

[2] H. Alhaija, S. Mustikovela, L. Mescheder, A. Geiger, and C. Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision*, 2018. 3

[3] G. Bertasius, J. Shi, and L. Torresani. Semantic segmentation with boundary neural fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 1, 3

[4] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3241–3248, 2010. 3

[5] D. Cheng, G. Meng, S. Xiang, and C. Pan. Fusionnet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(12):5769–5783, 2017. 3

[6] 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 *IEEE conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 1, 3

[7] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan. What is a good evaluation measure for semantic segmentation?. In *British Machine Vision Conference*, 2013. 2, 3

[8] L. Deng, M. Yang, H. Li, T. Li, B. Hu, and C. Wang. Restricted deformable convolution-based road scene semantic segmentation using surround view cameras. *IEEE Transactions on Intelligent Transportation Systems*, 2019. 1

[9] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 1, 3

[10] R. Fabbri, L. D. F. Costa, J. C. Torelli, and O. M. Bruno. 2d euclidean distance transform algorithms: A comparative survey. *ACM Computing Surveys (CSUR)*, 40(1):1–44, 2008. 4

[11] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. 3

[12] E. Fernandez-Moral, R. Martins, D. Wolf, and P. Rives. A new metric for evaluating semantic segmentation: leveraging global and contour accuracy. In *IEEE Intelligent Vehicles Symposium*, 2018. 2, 3

[13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1, 3, 5

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 2

[15] M. H. Hesamian, W. Jia, X. He, and P. Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019. 1

[16] R. A. Horn. The hadamard product. In *Proc. Symp. Appl. Math*, volume 40, pages 87–169, 1990. 5

[17] B. Kayalibay, G. Jensen, and P. van der Smagt. Cnn-based segmentation of medical imaging data. *arXiv preprint arXiv:1701.03056*, 2017. 1

[18] L. Ladickỳ, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In *IEEE International Conference on Computer Vision*, pages 739–746, 2009. 3

[19] S. Liu, W. Ding, C. Liu, Y. Liu, Y. Wang, and H. Li. Ern: Edge loss reinforced semantic segmentation network for remote sensing images. *Remote Sensing*, 10(9):1339, 2018. 3

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 2

[21] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690, 2014. 1

[22] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135:158–172, 2018. 1, 3

[23] 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 & Machine Intelligence (TPAMI)*, (5):530–549, 2004. 3, 5

[24] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[25] J. B. Roerdink and A. Meijster. The watershed transform: Definitions, algorithms and parallelization strategies. *Fundamenta Informaticae*, 41(1, 2):187–228, 2000. 3

[26] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European Conference on Computer Vision*, pages 1–15. Springer, 2006. 3

[27] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5229–5238, 2019. 1, 2, 3

[28] M. Thoma. A survey of semantic segmentation. *arXiv preprint arXiv:1602.06541*, 2016. 3

[29] Y. Yuan, J. Xie, X. Chen, and J. Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *European Conference on Computer Vision*, pages 489–506. Springer, 2020. 3

[30] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 3

[31] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 3