# Efficient Semantic Segmentation for Autonomous Vehicles with Limited Computational Resources
## [Evaluation Metrics]

Felix Blanco, *Member, UTEC*

## 1 Introduction

Semantic segmentation evaluation metrics are used to measure the performance of a semantic segmentation model, accurately segmenting the pixels of an image or classifying each pixel into a semantic category/region. This metrics provides quantitative measures to compare how model's predictions match the ground truth annotations [1], [2].

## 2 Evaluation Metrics

### 2.1 Pixel Accuracy

*Pixel Accuracy (PA)* is the simplest metric used to evaluate the performance of a semantic segmentation model. It is calculated by cumputing the ratio between the number of correctly classified pixels and the total number of pixels in the image [3]. The pixel accuracy is a value between 0 and 1, with 1 being the best possible score.

$$PA = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{1}$$

where $TP_i$ is the number of true positives (correct *foreground pixels*), $TN_i$ is the number of true negatives (correct *background pixels*) meaning correctly classified pixels, $FP_i$ is the number of false positives (incorrect *foreground pixels*) and $FN_i$ is the number of false negatives (incorrect *background pixels*) meaning incorrectly classified pixels.

Although the pixel accuracy is a simple metric, it is not a good metric to evaluate the performance of a semantic segmentation model. This is because the pixel accuracy does not take into account the *class imbalance problem*, presented in semantic segmentation datasets and real world scenarios.

### 2.2 Mean Pixel Accuracy

*Mean Pixel Accuracy (mPA)* calculate the pixel accuracy($PA$) for each semantic category and then averages the results [3]. The mPA is a value between 0 and 1, with 1 being the best possible score.

$$mPA = \frac{1}{n}\sum_{i=1}^{n} PA_i \tag{2}$$

using the same variables as in equation 1 and $n$ is the number of semantic categories. However, the mPA still does not take into account the *class imbalance problem*.

### 2.3 Intersection over Union

*Intersection over Union (IoU)* or *Jaccard index* is calculated by dividing the intersection of the predicted segmentation and the ground truth segmentation by the union of the predicted segmentation and the ground truth segmentation [3], [4]. The IoU is a value between 0 and 1, with 1 being the best possible score.

$$IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{TP_i}{TP_i + FP_i + FN_i} \tag{3}$$

where $A$ is the predicted segmentation, $B$ is the ground truth segmentation, $|A \cap B|$ represents the *overlapping area* between $A$ and $B$ meaning correctly classified pixels as foreground. While $|A \cup B|$ represents the *union area* between $A$ and $B$ meaning all pixels in foreground.

It is a useful metric to evaluate the performance of a semantic segmentation when the *class imbalance problem* is present, take into account the presence of small objects and penalize false positives. However, the IoU is still not a good metric alone to measure the performance of instance segmentation models, because it does not evaluate the background accuracy.

### 2.4 Mean Intersection over Union

*Mean Intersection over Union (mIoU)* calculate the IoU for each semantic category and then averages the results [3], [4]. The mIoU is a value between 0 and 1, with 1 being the best possible score.

$$mIoU = \frac{1}{n}\sum_{i=1}^{n} IoU_i \tag{4}$$

using the same variables as in equation 3 and $n$ is the number of semantic categories. The mIoU is a good metric to evaluate the performance of a semantic segmentation model, but it still does not evaluate the background accuracy.

### 2.5 Instance-level Intersection over Union

*Instance-level Intersection over Union (iIoU)* is calculated at the level of each instance, by dividing the intersection of the predicted segmentation and the ground truth segmentation by the union of the predicted segmentation and the ground truth segmentation [2], [4]. The iIoU is a value between 0 and 1, with 1 being the best possible score.

$$iIoU = \frac{|A \cap B|}{|A \cup B|} = \frac{iTP_i}{iTP_i + FP_i + iFN_i} \tag{5}$$

where $A$ is the predicted segmentation, $B$ is the ground truth segmentation, also using $iTP_i$ and $iFN_i$ to represent the number of true positives and false negatives at the level of each instance. They are computed by weighting the contribution of each pixel by the ratio class average instance size to the size of the instance [4].

This metric is a good metric to evaluate the performance of instance segmentation models, when large objects are present in the dataset.

### 2.6 Frequency Weighted Intersection over Union

*Frequency Weighted Intersection over Union (FWIoU)* extends the mIoU metric by taking into account the *class imbalance problem* by assigning a weight to each semantic category based on the number of pixels in the ground truth. This means that the FWIoU penalizes more the incorrect classification of pixels in semantic categories with more pixels in the ground truth, providing a more balanced evaluation [3], [5]. The FWIoU is a value between 0 and 1, with 1 being the best possible score.

$$FWIoU = \frac{1}{\sum_{i=1}^{n} |B_i|} \sum_{i=1}^{n} |B_i| IoU_i \tag{6}$$

where $|B_i|$ is the number of pixels in the ground truth for the semantic category $i$ (*frequency*) and $n$ is the number of semantic categories.

## 3 Paper Evaluation Metrics

ENet [4] uses the IoU metric[3], mIoU metric[4] and iIoU metric[5] to evaluate the performance of the model using the CamVid and Cityscapes datasets [6] against SegNet [7].

ENet uses both level of granularity, *class* and *category*, to evaluate the performance of the model. The $IoU_{class}$ and $IoU_{category}$ by using the predicted mask for specific class/-category and the ground truth mask for same class/category as before in both operations: $A \cap B$ and $A \cup B$. The mIoU uses the class/category IoU to calculate the mean IoU for each class/category. Also uses $iIoU_{class}$ and $iIoU_{category}$ to evaluate the performance of the model at instance level.

This gives a more detailed evaluation of the model's performance, at more granular level, than just using the mIoU metric.

## References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[2] A. Intelligence, "A modern approach, 4th us ed," *Retrieved October*, vol. 29, 2021.

[3] 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.

[4] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[5] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.