BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Center, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2024, Volume 15, Issue 4, pages: 295-318.
Submitted: August 7th, 2024 | *Accepted for publication:* November 19th, 2024

Selecting the Right Metric: A Detailed Study on Image Segmentation Evaluation

Giorgiana Violeta Vlăsceanu
Computer Science and Engineering Department, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania
giorgiana.vlasceanu@upb.ro
https://orcid.org/0000-0003-1664-3812

Nicolae Tarbă
Computer Science and Engineering Department, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania
nicolae.tarba@upb.ro
https://orcid.org/0000-0002-7769-8289

Mihai-Lucian Voncilă
Computer Science and Engineering Department, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania
mihai_lucian.voncila@stud.acs.upb.ro
https://orcid.org/0000-0002-6764-9125

Costin Anton Boiangiu
Computer Science and Engineering Department, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania
costin.boiangiu@cs.pub.ro
https://orcid.org/0000-0002-2987-4022

Abstract: *Image segmentation is critical role in various computer vision tasks, such as object recognition, medical imaging, autonomous driving, and document analysis. To evaluate the performance of segmentation algorithms, various metrics have been developed, each providing insights into different aspects of segmentation accuracy, object completeness, and boundary precision. This paper presents a comprehensive review of the key evaluation metrics used in image segmentation, including popular metrics such as Intersection over Union, Dice Coefficient, F1-score, and boundary-based metrics like average and maximum boundary distance. We explore the taxonomy of evaluation techniques, distinguishing between supervised and unsupervised methods, and discuss the strengths, limitations, and applications of each metric. Furthermore, we address the challenges associated with image segmentation, including handling noise, occlusions, illumination variation, and the scarcity of annotated datasets for training. By offering a detailed analysis of evaluation metrics, this paper aims to guide researchers and practitioners in selecting appropriate metrics for specific tasks and improving algorithm performance through informed comparisons and parameter tuning.*

Keywords: *Image Segmentation; Segmentation Evaluation Metrics; Segmentation Performance; Unsupervised Evaluation.*

How to cite: Vlăsceanu, G. V., Tarbă, N., Voncilă, M. L., & Boiangiu, C. A., (2024). *Selecting the right metric: A detailed study on image segmentation evaluation. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 15*(4), 295-318. https://doi.org/10.70594/brain/15.4/20

1. Introduction

Image segmentation is an essential process in computer vision in which an image is partitioned into multiple regions and segments, with each segment corresponding to a distinct object or region of interest. It is essential to many applications, including autonomous driving, medical imaging, scene comprehension, object recognition, and more. By segmenting images, valuable information can be extracted, allowing for targeted analysis and facilitating higher-level computer vision tasks. In the domain of document segmentation, this process involves partitioning a document image into meaningful regions, such as text, images, tables, headers, footers, and other structural components. Segmenting document images makes it possible to extract and analyze individual elements, enabling tasks like text recognition, document understanding, and layout analysis.

These are just a couple of examples of the many algorithms available for image segmentation. Some notable methods include U-Net (Ronneberger et al., 2015) used in biomedical image analysis, GrabCut (Rother et al., 2004) an iterative graph-cut-based algorithm that combines interactive user input and graph optimization, Random Walker (Xia et al., 2020), and Mask R-CNN (He et al., 2017). The choice of algorithm depends on factors such as the specific segmentation task, image characteristics, computational efficiency, and the availability of labeled training data.

1.1. Importance of Evaluation Metrics

Evaluation metrics play a crucial role in image segmentation, as they provide a quantitative and objective assessment of the performance and quality of segmentation algorithms. They are essential for comparing different segmentation methods, fine-tuning algorithm parameters, and measuring progress in the field of computer vision.

Evaluation metrics facilitate performance comparison by enabling the comparison of several segmentation methods or variants of the same technique. By quantitatively assessing metrics such as accuracy, precision, recall, or Intersection over Union (IoU), it becomes possible to identify which algorithms or techniques yield superior segmentation results. This helps in selecting the most suitable algorithm for a particular task or dataset.

Another reason is related to algorithm selection. Evaluation metrics make it possible to choose a segmentation method objectively that is suitable for a given application or dataset. Different segmentation methods have varying strengths and limitations, and evaluation metrics help identify algorithms that perform well in terms of desired criteria. If boundary accuracy is critical, measurements such as the mean absolute distance to ground truth can be employed to assess and identify the most precise algorithm.

Also, parameter tuning can be chosen based on evaluation metrics. Many segmentation algorithms have parameters that need to be configured for optimal performance. Evaluation metrics provide a means to quantitatively assess the impact of different parameter settings on segmentation results. By systematically varying these parameters and evaluating the corresponding metrics, researchers can determine the optimal parameter values that yield the best segmentation performance.

Nonetheless, evaluation metrics provide a standardised way to benchmark segmentation algorithms and track progress in the field. Datasets with ground truth annotations and established evaluation protocols allow researchers to compare their methods against approaches presented in the literature. This helps in measuring advancements in segmentation algorithms over time and enables the identification of novel techniques that outperform existing ones.

Finally, evaluation metrics facilitate dataset analysis by quantifying the quality and consistency of annotations. They help identify challenging image regions, cases of mislabeling or ambiguity, and potential biases in the dataset. Understanding dataset characteristics through evaluation metrics can guide the development of better segmentation algorithms and highlight areas for dataset improvement.

Image segmentation evaluation measures often utilised include IoU, pixel accuracy, Dice coefficient, precision, recall, F1-score, and boundary-based metrics like average or maximum boundary distance, or V-measure. These metrics provide insights into segmentation accuracy, object completeness, boundary localization, and overall performance.

Overall, evaluation metrics are vital for objective and standardised assessment of segmentation algorithms, aiding in performance comparison, algorithm selection, parameter tuning, benchmarking, and dataset analysis. They help advance the domain by promoting the development of more accurate and robust algorithms.

In the following sections, we will explore and discuss various image segmentation algorithms, their underlying principles, advantages, limitations, and real-world applications. By understanding the capabilities and trade-offs of different algorithms, researchers and practitioners could make knowledgeable choices when deciding a suitable segmentation method for their requirements.

1.2. Taxonomy on Evaluation of Image Segmentation

Zhang (1996) proposed three groups to classify the evaluation techniques of image binarization: analytical, empirical, and empirical discrepancy methods. When he refers to analytical methods, he refers to accessing the quality of the image segmentation by analysis of the mathematical proprieties. Segmentation algorithm authors often offer analytical assessments in the shape of proofs proving that the method improves specific criteria. Analytical evaluation, in contrast to other methods of assessment, is based on the mathematical characteristics of the algorithm instead of the characteristics of its output. This type of assessment cannot be automated, as it requires a deep understanding and analysis of the algorithm's mathematical foundations.

The empirical group detected by Zhang are methods that depend on the results of the segmentation algorithms. The last category - empirical discrepancy methods, involves calculating a measure of uniformity along the output of the segmentation and a ground truth segmentation. Note that the empirical method doesn't need a ground truth reference, instead, the performance is defined by analyzing specific properties of the output that are considered eligible.

Jiang et al. (2006) came up with a two-categories classification, in which we have the theoretical evaluation and the experimental evaluation. Zhang's analytical section is consistent with the theoretical evaluation, and Zhang's empirical category is identified by the experimental evaluation.

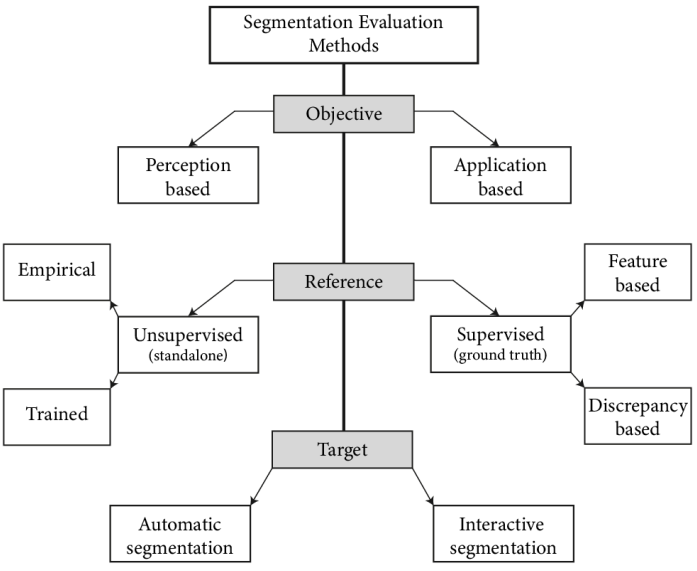

Figure 1. Taxonomy of segmentation evaluation algorithms from McGuinness(2009)

A series of articles use the term supervised for the method that utilised the ground truth (Yang et al., 1995) (Chabrier et al., 2006) (Zhang et al., 2008) and for those that do not - unsupervised. Another approach is presented by Correia and Pereira (2003) and introduces the terms standalone and relative for the same methods presented before.

McGuinness (2009) presented an interesting overview of the taxonomy of metrics for evaluation, which can be seen in Figure 1. In Section 4, we present a series of metrics based on the reference criteria, which characterise the methods that are using ground truth or not.

2. Overview of Image Segmentation

2.1. The Meaning and Applications of Image Segmentation

The process of dividing a picture into several logical and significant segments, each of which represents a different object or area of interest, is known as image segmentation. It is a basic computer vision task that is essential to many applications.

Image segmentation enables targeted analysis and understanding of images by extracting and isolating specific regions, objects, or structures. It allows for more accurate object recognition, tracking, and measurement, as well as facilitating higher-level tasks such as scene understanding, medical image analysis, autonomous driving, and content-based image retrieval. By segmenting a picture into areas that have semantic significance, image segmentation provides a powerful means of data representation, enabling efficient processing and facilitating the interpretation and manipulation of visual information.

2.2. Typical Limitations and Problems with Image Segmentation

Image segmentation is a complex task that faces several challenges and issues. These challenges arise due to the variability and complexity of image content, noise, occlusions, varying lighting conditions, and the diverse characteristics of objects and regions.

Images often contain regions that are visually similar or overlap, making it difficult to separate them accurately. This ambiguity poses challenges to assign correctly pixels to their respective segments and can result in segmentation errors.

Object occlusion represents another challenge. Occlusion occurs when objects or regions of interest are partially or completely obscured by other objects or the background. Handling occlusions during segmentation is challenging since it requires inferring the presence and boundaries of occluded objects based on available visual cues. In this list, object shape variation can also be added. Objects in images can exhibit significant shape variations, leading to difficulties in segmenting objects with complex and irregular boundaries. Accurately capturing the shape and contours of objects under different perspectives, scales, and deformations is a challenging task.

Image noise, caused by sensor limitations, compression artifacts, or environmental factors, can interfere with segmentation algorithms. Noise reduction techniques may be required to enhance the quality of input images and reduce false segmentation caused by noisy pixels.

Another factor worth mentioning is illumination variation. Inconsistent lighting conditions across an image or variations in illumination can pose challenges for segmentation algorithms. Shadows, highlights, and uneven illumination can affect the appearance and contrast of objects, making their segmentation more challenging. Illumination normalization techniques may be necessary to address these variations.

In the case of machine learning algorithms, a challenge is limited training data. Many segmentation algorithms rely on supervised learning approaches that require annotated training data. However, acquiring large-scale, high-quality annotated datasets for every segmentation task is time-consuming and expensive. Limited training data can impact the generalization and accuracy of segmentation algorithms. Some segmentation algorithms, especially those based on advanced techniques like deep learning, can be computationally intensive and require significant

computational resources. Real-time segmentation applications or resource-constrained environments may require efficient and lightweight segmentation algorithms.

Last, but not least, a common factor is user interaction and subjectivity. Interactive segmentation methods often involve user input or guidance to refine segmentation results. Subjectivity in user-defined regions or constraints can introduce biases and affect the reproducibility and objectivity of segmentation algorithms.

To deal with these challenges, it is essential to develop robust algorithms capable of managing diverse visual content, that incorporate contextual information, adapt to evolving conditions, and effectively deal with noise and uncertainty. Ongoing research aims to improve the reliability, efficiency, and accuracy of image segmentation techniques, in response to these issues.

2.3. Current Image Segmentation Algorithms and Solutions

Image segmentation can have two categories of approaches: non-contextual, when the algorithm analyzes each pixel, and contextual, which takes into consideration the relationship between pixels (which might be location, color, or similar intensity). Depending on the method employed by an algorithm to create distinct groups, image segmentation can be based on layers for solutions based on CNNs or blocks of pixels such as regions and edges.

Image segmentation methods that utilise regions are based on identifying similarities in the image. Some examples for this category are clustering algorithms such as k-means, algorithms based on histograms or different thresholding values, watershed, compression, graphs, or mean-shift.

Edge detection is a subset of image segmentation techniques. When we are discussing segmentation for edge detection, these methods are based on determining different discontinuities in the image. Edge detection techniques are grouped based on the operator used in the process of detecting the edge, such as Canny (1986), Sobel (Pinastawa et al., 2024), Laplacian (Wang et al., 2007), and Prewitt (Prewitt et al., 1966).

2.4. Evaluation Metrics' Significance in Image Segmentation

Evaluation metrics have a broad significance in a variety of fields related to image segmentation, as they provide a quantitative and objective assessment of the performance and quality of segmentation algorithms. They are essential for comparing different segmentation methods, fine-tuning algorithm parameters, and measuring progress in the field of computer vision.

For instance, in the case of object detection and recognition, which can be used in tasks such as autonomous driving to separate pedestrians and other vehicles from the background, metrics such as IoU are heavily relied upon in order to determine how accurately objects are detected and delineated in an image.

Similarly, in the case of healthcare, segmentation is used to identify specific regions in medical images, such as those pertaining to tumors. In such a case metrics like the Dice Coefficient or specificity and sensitivity help in evaluating whether the precision of segmenting diseased and healthy tissues is appropriate since it can affect overall diagnosis and patient outcomes.

Another field where selecting the right evaluation metric where segmentation is crucial could be that of document retrieval and restoration. In these contexts, there is a need for a strong separation between foreground and background in order to accurately retrieve the image text. Thus, metrics that can assess the accuracy and precision of such boundaries are extremely important, since they ensure minimal artifacts and better visual results.

Lastly, in the case of deep learning models and other techniques based on supervised training, it is important to evaluate the consistency and accuracy of annotations across specific datasets. This allows for higher-quality training data.

Overall, evaluation metrics help advance the field of image segmentation by promoting the development of more accurate and robust algorithms.

3. Evaluation Metrics for Image Segmentation

Metrics in general are used to measure and evaluate the performance of an algorithm or set of steps. Evaluation metrics enable researchers to compare the performance of several segmentation algorithms or variants of the same algorithm when it comes to image segmentation. By quantitatively assessing metrics such as accuracy, precision, recall, or IoU, it becomes possible to identify which algorithms or techniques yield superior segmentation results. This helps in selecting the most suitable algorithm for a particular task or dataset.

Evaluation metrics also make it possible to choose an objectively suitable segmentation algorithm for a certain dataset or application. Different segmentation methods have varying strengths and limitations, and evaluation metrics help in identifying algorithms that perform well in terms of desired criteria. Metrics like the mean absolute distance to ground truth, for instance, can be used to assess and choose the most accurate method where boundary precision is crucial.

In the same light, evaluation metrics can improve parameter tuning. Many segmentation algorithms have parameters that need to be configured for optimal performance. Evaluation metrics provide a means to quantitatively assess the impact of different parameter settings on segmentation results. By systematically varying these parameters and evaluating the corresponding metrics, researchers can determine the optimal parameter values that yield the best segmentation performance. Also, metrics provide a standardised way to benchmark segmentation algorithms and track the progress in the field. Datasets with ground truth annotations and established evaluation protocols allow researchers to compare their methods against well-known approaches from the literature. This helps measure advancements in segmentation algorithms over time and enables the identification of novel techniques that outperform existing ones.

Last, but not least, evaluation metrics facilitate dataset analysis by quantifying the quality and consistency of annotations. They help identify challenging image regions, cases of mislabeling or ambiguity, and potential biases in the dataset. Understanding dataset characteristics through evaluation metrics can guide the development of better segmentation algorithms and highlight areas for dataset improvement.

In the following sections, we explore the metrics associated with image segmentation, using the taxonomy pointed out before.

4. Supervised Evaluation

This type of evaluation targets segmentation accuracy in terms of the number of regions determined, their location, and their computed size. The cumulative difference or discrepancy between matching regions in an image can be used to describe a region-based comparison of two segmented images.

4.1. Precision, Sensitivity (or Recall) and Specificity

The basic metrics for binary classification are specificity, sensitivity (or recall), and precision. When combined, they offer a thorough understanding of an algorithm's capabilities (Yang et al., 2023) (Ntirogiannis et al., 2013) (Yap et al., 2004).

Precision

The following formula is used to compute precision:

$$Precision = \frac{TP}{TP + FP} \qquad (1)$$

This statistic represents the percentage of all algorithmic positive predictions that are true positives (pixels correctly identified in a segment).

Precision's equation (TP) stands for true positives and (FP) false positives (pixels were mistakenly recognised in a segment). Increased precision lowers false alarms by demonstrating the algorithm's dependability in identifying a pixel being in the segment.

Sensitivity

This measure expresses the percentage of true positive cases that the algorithm was able to correctly identify. It is defined mathematically as:

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

False negatives, or pixels mistakenly labeled in a group, are indicated by (FN). A higher sensitivity indicates that the algorithm is more adept at identifying every pixel in a group, which lowers the number of missed detections (Ntirogiannis et al., 2013).

Specificity

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

This measure shows the percentage of true negative instances (pixels not part of the segment) that the algorithm successfully detects. The calculation is done using Equation 3. As is shown in the Specificity's equation, (TN) indicates true negatives. According to Yap et al. (2004), a method with a higher Specificity produces fewer errors in identifying pixels that are not in the segment.

By weighing an algorithm's accuracy for both categories of pixels and considering its tendency for false negatives (Sensitivity) and false positives (Precision), these metrics together offer an in-depth assessment of the algorithm's performance.

Unfortunately, just a single part of the performance is captured by each of these measurements, which could result in a poor representation. For example, an algorithm may be excessively conservative when classifying the pixels in an image, if it has high Precision but poor Sensitivity, or the opposite (Ntirogiannis et al., 2013) (Yap et al., 2004). More proficient metrics, such as the Distance-Reciprocal Distortion Measure (DRD) can provide a deeper perspective on the evaluation.

However, when assessing classification techniques, Precision, Sensitivity, and Specificity are usually utilised as starting points (Ntirogiannis et al., 2013) (Yap et al., 2004) because of their interpretability and simplicity.

4.2. F-measure

Recall and precision are combined in the F-measure, also referred to as the F-score or F1-score (Ntirogiannis et al., 2013). The F-measure attributes a relationship in the form of a harmonic mean between the two, as can be observed in Equation 4.

$$F_1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

If we want a broader perspective, we can take a look at F_β, a more comprehensive F-score, which employs a positive real factor β where β is selected so that recall is deemed β times as significant as precision. It looks like this:

$$F_\beta = \left(1 + \beta^2\right)\frac{Precision \cdot Recall}{\left(\beta^2 \cdot Precision\right) + Recall} \quad (5)$$

It can be observed that in the case of $\beta = 1$, we obtain the F1-score. What's interesting about this metric is the use of the harmonic mean, which imposes a penalty on its components. Because of this, the F1-score is more sensitive to the lower values of precision and recall, making the overall value closer to the minimum of these two metrics, as opposed to other similar averages, such as arithmetic mean and geometric mean.

The F-measure can be interpreted fairly in simple terms: a higher F-measure denotes superior segmentation algorithm performance. Perfect recall and precision are indicated by an

F-measure of 1, while total failure in both areas is indicated by an F-measure of 0 (Ntirogiannis et al., 2013).

The main benefit of the F-measure is its capacity to offer a fair assessment of an algorithm's efficiency. It makes sure that neither recall nor precision is overemphasised, reducing the possibility of an assessment that is either optimistic or pessimistic.

On the other hand, the F-measure tends to have certain issues, more specifically assuming that recall and precision are equally important, which, in some circumstances, might not be the case, such as those where datasets might contain imbalanced classes. For instance, in the case of medical images, recall might be more desirable than precision, as missing tumors can be more detrimental than misidentifying healthy tissue. In this scenario, sometimes the F2 is used as opposed to the F1 since it puts a higher emphasis on identifying FN. Hashemi et al. (2018) use this in their study alongside a wide array of other metrics such as the Dice Index and Jaccard Index to evaluate their proposed model used for segmenting images that contain multiple sclerosis.

Furthermore, similarly to the F-measure, the F_{β}, regardless of the value β, provides an overall assessment and may not sufficiently reflect structural aspects or local distortions (Bharathi et al., 2019). Metrics such as the Distance-Reciprocal Distortion Measure (DRD) and Pseudo-F-measures were initiated in some contests in response to these restrictions (Bharathi et al., 2019).

Despite these limitations, the interpretability and balance it provides between recall and precision make the F-measure a popular metric in classifying problems in image processing.

4.3. Dice Coefficient

Also named the Sørensen-Dice index, the Dice coefficient (DSC) is a statistic for comparing the resemblance of two samples. In the context of image segmentation, it is employed to look at the degree of agreement between the predicted segmentation and the ground truth (Setiawan et al., 2020). By dividing the area where the ground truth and the predicted set intersect by the sum of their respective sizes, one can get the DSC. Regarding the mathematical formula, it might be stated as follows:

$$DSC = \frac{2TP}{2TP+FP+FN} \qquad (6)$$

In essence, DSC indicates the accuracy of the segmentation algorithm while being sensitive to overlap, since it puts a higher emphasis on TP, and gives them twice the weight in the calculation, as opposed to incorrectly identified pixels, which are less penalised. Something to note is that the DSC is both balanced and symmetrical. It is balanced because it performs the same for both over- or under-segmentation since it puts the same emphasis on FP and FN. It is symmetrical because it doesn't matter if we consider the ground truth image as the reference set or the predicted image as the reference set, the result will be the same.

The enhanced performance of the method is indicated by a DSC that is closer to 1, which indicates a high degree of similarity between the segmented image and the ground truth. One of the DSC's advantages is that it is simple to calculate and understand. It provides a fair assessment of the efficacy of the segmentation method by accounting for both false negatives and false positives.

The DSC can be impacted by class imbalances in the image, just like any other ratio-based metric, such as the F1-score, making it sensitive to the distribution of positive and negative classes. To provide a more comprehensive assessment, it can be used in conjunction with other metrics, such as Pseudo-F-Measure, Jaccard Index, or DRD to help balance precision, recall, and other aspects related to accuracy. Despite its sensitivity to class imbalance, it remains an important metric for evaluating image segmentation methods.

4.4. Jaccard Index

A metric for comparing sample set diversity and similarity is the Jaccard similarity coefficient, also known as the Jaccard Index or Intersection over Union (IoU) (Ogwok et al., 2022).

Mathematically speaking, the Jaccard Index is equal to the intersection's size divided by the dimension of the two groups' union and for a classification task can be formulated as Equation 7.

$$IoU = \frac{TP}{TP+FP+FN} \qquad (7)$$

In other words, the Jaccard Index compares the ground truth segmentation with predicted segmentation and gives a degree of overlap between the two regions. Values with greater segmentation are closer to one; those with worse segmentation are farther from one or closer to zero (Ogwok et al., 2022).

The Jaccard Index does, however, have some drawbacks. It might ignore local distortions in the image because it is a global metric. Additionally, it may be affected by the image's class balance, just like any other ratio-based statistic.

Despite these drawbacks, the Jaccard Index is a valuable tool that enhances other measures, such as the F-measure, Pseudo-F-measure, and (DRD), offering a thorough assessment of the effectiveness of image segmentation methods.

According to (Bertels et al., 2019), the Jaccard index and the Dice coefficient are equivalent in applications with high-class imbalances, such as medical image segmentation, where the positive class is by far outnumbered by the negative class.

4.5. Tversky Index

One of the main issues of the DSC is that it equally balances the FP and FN in an image. In order to address this class imbalance, the Tversky Index (TI), described by Abraham & Khan (2019) introduces two new weighting parameters α, β.

$$TI = \frac{TP}{TP+\alpha FP+\beta FN} \qquad (8)$$

One thing to note is that $\alpha, \beta = 0.5$ the TI becomes the DSC, whereas if $\alpha = 1, \beta = 0$, the coefficient becomes the precision, in the opposite case, it becomes the recall. This flexibility in choosing which information to prioritise tends to be particularly useful in the field of medical image segmentation, where it aims at both dealing with class imbalances, as well as fine-tuning precision and recall.

In the context of class imbalances, medical images tend to have tumors, or lesions (foreground), that are very often much smaller than the tissue around them (background). In this case, the DSC or the Jaccard Index will not work well since they will favor the background class. TI would allow for more precise control in giving more weight to the FN than to the FP, thus prioritizing recall over precision, as missing a tumor could be much worse than misclassifying a healthy tissue.

In a similar manner, in the case where you would try to identify blood vessels, you would try to place a higher emphasis on precision rather than recall, in order to avoid misclassifying healthy tissue as blood vessels.

For these reasons, the TI tends to be ideal for a wide range of medical applications. (Salehi et al., 2019) use a loss metric based on TI for training fully convolutional deep networks to segment medical images containing lesions, which occupy a very small part of the image, and obtain a good trade-off between precision and recall.

4.6. Average Precision and Mean Average Precision

A metric used to evaluate the trade-off between precision and recall is the one of Average Precision. In order to compute the value of this metric a precision-recall curve is created, by computing the precision and recall at various confidence score thresholds and plotting it using the precision as the y-axis and the recall as the x-axis. Then the area is computed using interpolation, as can be seen in Equation 9.

$$AP = \frac{1}{n}\sum_{i=1}^{n} P\left(r_i\right) \tag{9}$$

Where $P(r_i)$ is the precision at the recall point r_i, and n is the number of recall points used. In practice, the most used method is that of 11-point interpolation where we can consider 11 points, evenly spaced out in the interval [0, 1], but for better accuracy, more points can be used, though this is typically not required.

In the case of multiple classes that need to be detected, we can compute the mean Average Precision (mAP) across all classes:

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{10}$$

Where C is the total number of classes, and AP_i. One thing to note is that for specific datasets such as COCO (Lin et al., 2015) the mAP is computed across multiple IoU thresholds in order to evaluate a classification model's robustness and ability to handle various overlap conditions. This is typically taken in the interval [0.5, 0.95] with a step of 0.05. Thus, a version which takes into account a specific threshold value such as 0.5 would be:

$$mAP@50 = \frac{1}{C}\sum_{i=1}^{C} AP_i^{IoU_{0.5}} \tag{11}$$

Where the $mAP@50$ denotes that the precision considers just objects that intersect for at least 50% of their area. Thus, we can compute the mAP for the entire dataset according to the number of increments taken:

$$mAP@[IoU = 50{:}95] = \frac{1}{C}\sum_{i=1}^{C} AP_i^{IoU_{50}} \tag{12}$$

4.7. Area Under the Curve and Receiver Operating Characteristics

When it comes to classification problems, Area Under the Curve (AUC) and ROC (Receiver Operating Characteristics) are handy for checking any classification model's efficiency. In contrast to ROC, which is a probability curve, AUC is a metric or degree of separability. It shows the extent to which the model is capable of class discrimination. The greater the AUC, the better the model predicts certain classes as their ground truth counterparts.

Sensitivity (Recall or TPR – True Positive Rate) is placed on the y-axis and FPR (False Positive Rate, which is $1 - Specificity$) is plotted on the x-axis to create the ROC curve.

For segmentation tasks, ROC analysis can be performed by treating each pixel as a binary classification problem (object vs. non-object). The curve reflects how well the segmentation model can differentiate between the classes (e.g., object vs. background) across varying threshold levels. A high AUC means the model effectively distinguishes between the segmented object and the background.

4.8. Matthews Correlation Coefficient

The phi coefficient or the Matthews Correlation Coefficient (MCC), is a classification metric that accounts for false and true positives as well as negatives. Even though there are significant size differences between the categories, it is frequently seen as a harmonic metric that may be applied (Itaya et al., 2024). The MCC is essentially a correlation coefficient between the binary classifications that were found and those that were predicted. A value in the range of -1 and +1 is returned. Here, a coefficient of +1 indicates an ideal forecast, a value of 0 indicates no better than a haphazard estimate, and a coefficient of -1 indicates the complete discrepancy between the forecast and the observation.

Based on the parts of the confusion matrix, MCC can be described as:

$$MCC = \frac{TP \bullet TN - FP \bullet FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (13)$$

The primary benefit of the MCC over other metrics is that it is a more dependable statistical rate that only yields a high score if the predictor performs well in each of the four confusion matrix parts, as opposed to accuracy, F-score, or the area under the ROC curve (AUC). Put differently, the MCC provides a fair assessment of the quality of classifications by taking into account both the over- and under-prediction of each class.

Nevertheless, a drawback of the MCC is that it lacks a clear interpretation in terms of odds ratios or estimates, and therefore does not translate well to multi-class classification. Furthermore, although it provides a more impartial viewpoint, it can be trickier to understand and articulate than more straightforward data like accuracy or F1 score (Itaya et al., 2024). Despite these limitations, the MCC is considered by many to be a good statistical accuracy measure when classes are unbalanced, and it serves as a useful addition to other evaluation metrics in image segmentation assessments.

4.9. Hamming Distance

The Hamming Distance was first introduced by Huang et al. (2002). They established correspondences between the regions in two segmentations, *S* and *R*, to maximise their overlap. The following represents the directional Hamming distance, from *S* to *R*:

$$D_H(SR) = \Sigma_{r_i \in R} \Sigma_{s_k \neq s_j, s_k \cap r_i \neq 0} \left| r_i \cap s_k \right| \qquad (14)$$

Hamming Distance is determined by summing the intersection areas between each region, r_i, in *R,* and the non-maximally intersected regions, s_j and s_k in *S*. The regions r_i in *R*, s_j, s_k in *S*, are assumed to intersect partially, but they do not fully overlap with one another. The total area of these intersections represents a measure of dissimilarity. However, if the regions r_i , s_j, and s_k do not intersect at all $\left(s_j \cap s_k \cap r_i = 0\right)$, then the dissimilarity measure will be zero.

To assess the regions based on this measure, a region-based evaluation metric - p is introduced using a normalised dissimilarity measure. It is computed as:

$$p = 1 - \frac{D_H(SR) + D_H(RS)}{2 \times |S|}, \qquad (15)$$

where $|S|$ represents the total number of elements in the region. The value p ranges from 0 to 1, where a smaller mismatch between the segmentation *S* and the reference region *R* results in a higher p value, approaching 1 when the two regions are very similar.

Hamming Distance is often used alongside other metrics like Dice Coefficient, and Intersection over Union (IoU), as it primarily captures the absolute error count but does not give insight into the spatial structure of the segmentation.

4.10. Local Consistency Error

Local Consistency Error (LCE) is introduced by Martin et al. (2002) as a metric to quantify the consistency between image segmentations with different levels of granularity. LCE enables the refinement of labeling between the segmentation and the ground truth, facilitating a more accurate assessment of their agreement.

$$LCE(S,R) = \frac{1}{N}\sum_{i} min\Big(E\Big(S,R,p_i\Big), E\Big(R,S,p_i\Big)\Big) \qquad (16)$$

The agreement between two segmentations S, R at a pixel p is quantified by $E(S,R,p)$, where E represents a measure of agreement. N represents the total number of pixels in the image.

The Local Consistency Error serves as an error measure, where no error is defined as a score of 0, and a 1 score corresponds to the maximum error. It is important to note that LCE is particularly meaningful when the two segmentations under comparison consist of a comparable quantity of segments. In Martin et al. (2002), the author pointed out that there are two specific scenarios where LCE yields a score of zero: when each segment consists of only one pixel, or when there is a single segment that encompasses the entire image.

LCE is a non-symmetrical measure. Martin et al. (2002) tried to rectify the downside of this metric by proposing Global Consistency Error (GCE) defined as:

$$GCE(S,R) = \frac{1}{N} min\left(\sum_{p_i} E\Big(S,R,p_i\Big), \sum_{p_i} E\Big(R,S,p_i\Big)\right) \qquad (17)$$

In terms of precision, LCE prioritises fine boundary precision, while GCE focuses on broader regional coherence. For tasks requiring high boundary fidelity (e.g., medical image segmentation), LCE may be more suitable. In contrast, GCE is more appropriate for tasks where general region consistency matters more than boundary precision.

LCE is sensitive to small, local errors, potentially penalizing a model even if it is globally accurate. This can lead to over-penalization in applications where minor boundary discrepancies are tolerable. GCE provides a more forgiving assessment by focusing on the global structure, potentially overlooking small but important details.

LCE's sensitivity to boundaries makes it suitable for tasks where detail along the edges is crucial, while GCE may better serve segmentation applications where the general shape and area coverage are more critical than edge accuracy.

4.11. Bidirectional Consistency Error

Monteiro and Campilho (2012) proposed an updated version of the LCE formula to address the issue of degenerate segmentations. He proposes a metric that penalises differences in segmentations in proportion to the extent of overlap between regions. A metric that does not allow for refining is produced by replacing the pixel-wise minimum with a maximum. This specific metric is known as the Bidirectional Consistency Error (BCE) and is defined as follows:

$$BCE(S,R) = \frac{1}{N}\sum_{i} max\Big(E\Big(S,R,p_i\Big), E\Big(R,S,p_i\Big)\Big) \qquad (18)$$

BCE demands high precision on both sides of the segmentation, which can be computationally expensive and sensitive to small errors in boundary areas. This can be challenging in cases with slight overlaps or minor boundary inconsistencies that would not significantly impact downstream tasks. BCE might lead to over-penalizing models in applications where small boundary deviations are acceptable, resulting in more complex model tuning.

This metric is primarily designed for binary or single-class segmentation. When multiple classes are involved, implementing BCE requires computing bidirectional consistency for each class independently, which can be less interpretable. Multi-class segmentation tasks (e.g., identifying multiple structures in a medical image) might need adapted versions of BCE or supplementary metrics like the Dice Coefficient or Intersection over Union (IoU).

4.12. Partition Distance Measure

Cardoso and Corte-Real (2005) demonstrated that a novel metric introduced by Almudevar and Field (1999), which they denoted as partition distance measure (PDM), d_{sym}, follows the properties of a distance. This is expressed as follows: "Given two partitions P and Q of S, the partition distance represents the minimum number of elements that need to be removed from S in order to make the induced partitions (P and Q restricted to the remaining elements) identical.", which can be rewritten to follow the formula in equation (14).

$$d_{sym}(P, Q) = |P \cup Q| - |P \cap Q| \qquad (19)$$

As we can notice, $d_{sym} = 0$ only if P = Q.

PDM is frequently used in clustering-based segmentation tasks, such as superpixel segmentation or k-means clustering, to evaluate how well clusters in the predicted segmentation align with clusters in the ground truth. This can be seen in medical imaging (e.g., organ segmentation) and document analysis (e.g., separating text from images or tables). This metric provides a quantitative assessment of clustering consistency, which is essential for applications where maintaining accurate clusters is critical, such as identifying distinct regions within a tissue sample.

In terms of trade-offs, PDM is sensitive to errors in the grouping, so it penalises both over-segmentation and under-segmentation. This is beneficial for applications where precise segmentation is essential (e.g. medical imaging).

Calculating PDM requires examining all pairs of pixels, which is computationally expansive for high-resolution images and large datasets. This limits its use in real-time applications or for high-resolution image segmentation.

4.13. Panoptic Quality

Kirilov et al. (2019) proposed a metric more suitable for panoptic segmentation which they denote PQ and calculate as the product of the average IoU of matched segments, which they refer to as Segmentation Quality (SQ), and the F1 score, which they refer to as Recognition Quality (RQ).

$$PQ = SQ \cdot RQ \qquad (20),$$

where $RQ = F_1$ and

$$SQ = \frac{\sum_{s \in S, r \in R,\ s \subseteq r \vee r \subseteq s} IoU(s,r)}{TP} \qquad (21).$$

PQ balances between object instances and overall background segmentation quality. However, this balance may not always reflect application priorities, especially when one aspect (like instance segmentation) is more important than the other.

PQ adds complexity in calculating FP and FN and penalises both FP and FN instances, so it can be sensitive to missed detections and extra instances.

4.14. Boundary-based Metrics

Boundary-based metrics assess the accuracy of boundary localization in segmented regions. These metrics measure the discrepancy between the segmented boundaries and the corresponding ground truth boundaries. Common boundary-based metrics include the average boundary distance

and the maximum boundary distance. The average Euclidean distance between each pixel on the segmented border and its nearest point on the ground truth border is determined via the average boundary distance calculation. The maximum Euclidean distance between any pixel on the segmented border and its nearest point on the ground truth boundary is measured by the maximum boundary distance (Bowen et al., 2021).

4.15. Average Boundary Distance

The average boundary distance measures the average Euclidean distance between each pixel on the segmented boundary and its nearest point on the ground truth boundary. It provides an estimate of the overall discrepancy between the boundaries. A smaller average boundary distance indicates a higher level of boundary localization accuracy.

$$AverageBoundaryDistance = \frac{1}{N}\sum dist(i), \qquad (22)$$

where N is the total number of boundary pixels and $dist(i)$ represents the Euclidean distance between the ith pixel on the segmented boundary and its nearest point along the limit of ground truth.

Average boundary distance is commonly used in evaluating the performance of image segmentation algorithms, particularly when boundary accuracy is of critical importance. It provides a quantitative measure of how well the algorithm captures the precise location of object boundaries.

4.16. Maximum Boundary Distance

The maximum boundary distance measures the maximum Euclidean distance between any pixel on the segmented boundary and its nearest point on the limit of ground truth. It represents the largest discrepancy between the boundaries and reflects the worst-case scenario in terms of boundary localization error.

$$MaximumBoundaryDistance = max(dist(i)), \qquad (23)$$

where $dist(i)$ represents the Euclidean distance between the ith pixel on the segmented boundary and the place on the ground truth boundary that is closest to it.

The maximum boundary distance is often used to assess the robustness of segmentation algorithms by identifying the most significant errors or outliers in boundary localization. It helps in understanding the worst-case performance of an algorithm and in detecting cases where the segmentation results deviate substantially from the ground truth.

These boundary-based metrics show how well segmentation algorithms capture object borders and offer insights into boundary localization accuracy.

4.17. Hausdorff distance

The Hausdorff distance is another metric that tends to be used in the context of evaluating segmentation quality when boundaries are of great importance. The metric compares two sets by calculating the maximum deviation between them, making it particularly sensitive to small errors in boundary alignment.

$$d_H = (||p - r||,\ ||p - r||\ ) \qquad (24)$$

Where P is the set of predicted regions, and R is the set of ground truth regions.

The metric is particularly useful in the context of medical image segmentation since it can delineate tumors or organs more easily and is sensitive to large misalignments between predicted and ground truth boundaries. However, it suffers from the issue that it tends to be overly sensitive to outliers, where a large discrepancy can affect the metric drastically, whilst also having a limited sensitivity to smaller errors, which might still be of significance.

A study that considers this metric is that of Hatamizadeh et al. (2019) which proposes an encoder-decoder architecture and evaluates it using the Dice Index, Jaccard Index, and Hausdorff distance.

4.18. Boundary F1-Score

In the context of boundary evaluation, a metric that is commonly used is that of the Boundary F1-Score (BF-Score). This metric aims at assessing image segmentation only at region boundaries as those tend to be more critical in certain domains, such as medical imaging. In such cases identifying the overall shape of an object is more important than identifying the whole object, as the overall size of the object can affect the final diagnosis.

The BF-Score considers boundary precision (Pr_B) and boundary recall (R_B), which consider only pixels at the boundaries of the predicted regions and the ground truth when computing TP, FP, FN. In order to determine if certain pixels should be matched, a distance d, that represents the threshold of what area would be considered a boundary, is used.

$$BF = 2\frac{Pr_B \bullet R_B}{Pr_B + R_B} \tag{25}$$

A recent study that takes advantage of this metric is one done by Li et al. (2024), which analyzes how their framework for Object-centric Temporal Action (OTAS) performs segmentation on various datasets related to cooking videos. In this study, they improve other state-of-the-art methods by 41% when computing using their proposed version of BF-Score. Though in their case the distance for threshold is considered from a time variation perspective, based on the length of the video, rather than the image size, using variations for 2% of the length for a BF small, and 5% for a BF large.

4.19. Boundary IoU

Another metric that uses the set of boundary pixels is that of the Boundary IoU (BIoU), proposed by Cheng et al. (2021). Similarly to the BF-Score, the BIoU uses a fixed distance threshold in order to ascertain which pixels sit on the boundary. The formula is similar to the classic IoU one, as can be seen in equation 26.

$$BIoU = \frac{B_p \cap B_{gt}}{B_p \cup B_{gt}} \tag{26}$$

Where B_p represent the boundary pixels in the prediction, and B_{gt} the boundary pixels in the ground truth.

This metric tends to be widely used in applications where boundaries are of critical importance such as medical imaging, or autonomous driving. A recent study that takes advantage of this metric is the one performed by Sun et al. (2023) which uses it both as an evaluation metric, as well as a modified loss function for three different architectures and tested on multiple different datasets.

Another study that uses this metric is the one performed by Lin et al. (2023), which creates an encoder-decoder structure for segmentation and uses boundary aids in order to improve the resulting segmentation.

4.20. Boundary AP

Also in recent years, studies have modified the use of AP, and mAP to consider the inclusion of the BIoU in their computations in order to increase the robustness of segmentations, such as medical imagining, Lin et al. (2023). The formulas remain unchanged in their way of usage, just that the IoU is replaced by BIoU.

However, there are cases outside of just medicine for such a metric. When trying to separate any object of interest from another that is close next to it, evaluating based on the boundary is always of interest. Such a case is when attempting to identify multiple animals that are close to each

other, such as sheep. Zhao et al. (2023) address this issue in their study by proposing SheepInt which is a framework used to segment sheep images, and they use boundary AP alongside other metrics to evaluate their results.

5. Unsupervised Evaluation

We presented, in the previous subsection, a series of evaluation metrics for image segmentation which operate by comparing the output with a ground truth image considered ideal. For this section, we discuss unsupervised evaluation metrics. In the absence of the ideal image, these metrics have the objective to evaluate the segmentation quality. By observing the output, without the input image, we can suspect when the algorithm has an over-segmented region.

We can affirm that this class of evaluation metrics focuses on broader characteristics of the segmentation, rather than assessing how well they describe a comparison to a reference image. Moreover, understanding the basis of human perceptual classification implies that the measures have this goal in the end. Our inclination when creating a method is to approach the issue from the bottom up, so in the end, the solution uses performant algorithms, and the effort is directed toward the greater result, and the quality of the segmentation comes later. The shift with this evaluation method is that it focuses exactly on what is implemented and gives a helping hand to maintain the attention on finding an algorithm that performs on specific criteria.

5.1. Entropy-based Metrics: V-measure

V-measure is an evaluation metric commonly used for assessing the quality of clustering or segmentation results. It provides a single score that balances both homogeneity and completeness of the clusters or segments. V-measure considers the agreement between the ground truth labels and the predicted labels, by considering the balance between homogeneity and completeness.

To define homogeneity, completeness, and V-measure respectively, we must first define the entropy and conditional entropy for two sets of segmentations. Entropy is defined as:

$$H(X) = -\sum_{x \in X} P(x) log(P(x)) \tag{27}$$

Where X represents the set of possible values, in our case the multitude of regions, and $P(x)$ represents the probability of each element belonging to a certain class.

Similarly, we can define the conditional entropy for two different sets, as follows:

$$H(X|Y) = -\sum_{y \in Y} P(y) \sum_{x \in X} P(x|y) log(P(x|y)) \tag{28}$$

Where X is the set representing the truth labels, and Y is the set of predictions, $P(y)$ represents the probability, and $P(x|y)$ represents the conditional probability.

Given these, we can now define homogeneity. Homogeneity measures the degree to which each segment contains only data points that belong to either a single class or the ground truth labels. In other words, it checks how consistent the predicted segments are with the actual ones. The formula for homogeneity can be seen in Equation 19, for a predicted region S, and a ground truth region R.

$$h = 1 - \frac{H(S)}{H(R)} \tag{29}$$

What is to note is that if the predicted regions are close to the ground truth regions, then the value of the conditional entropy, $H(S)$, will decrease towards 0, and the value h will increase towards 1. In the case where the predictions perfectly match the ground truth regions, $H(S) = 0$, and $h = 1$, indicating the segmentation is perfectly homogeneous.

Similarly, we can define the completeness. Completeness measures the degree to which all elements of a given class are grouped, in other words, it determines how well the segmentation captures entire regions.

$$c = 1 - \frac{H(R)}{H(S)} \quad (30)$$

As in the case of homogeneity, if the majority of the elements of a region are predicted correctly, then the conditional entropy, $H(R)$, will decrease towards 0, and the value c will increase toward 1. A value of $c = 1$ is achieved, only if all the predictions match the ground truth segmentations.

Based on these definitions for homogeneity and completeness, we can now define the V-measure as a fusion of these two metrics, calculated by determining their harmonic mean (Rosenberg et al., 2007):

$$V = \frac{2 \times h \times c}{h + c} \quad (31)$$

In terms of advantages, the V-measure provides a balanced measure by considering both homogeneity and completeness in its calculation, providing a comprehensive evaluation of clustering or segmentation quality. It addresses situations involving unique cluster sizes, taking into account the relative proportions of samples assigned to each cluster or segment, which is essential in handling imbalanced class distributions.

Also, it handles unequal cluster sizes. V-measure can handle cases where the clusters or segments have significantly different sizes or imbalanced class distributions. It considers the relative proportions of samples assigned to each cluster or segment.

However, it does have its limitations, particularly in terms of sensitivity to the number of clusters. As the number of clusters increases, the V-measure tends to favor solutions with more clusters. This can be problematic in cases where the ground truth has a different number of clusters than the predicted labels.

In conclusion, the V-measure is a useful evaluation metric for assessing the quality of clustering or segmentation results. It balances the homogeneity and completeness of the clusters or segments, providing a single score. However, it is important to consider its sensitivity to the number of clusters when interpreting the results.

5.2. Visible Color Distance Based Evaluation

Another approach for an unsupervised evaluation measure is grounded on color distances, more exactly on the perception of the colors. Chen & Wang (2004) proposed two measures to measure the perception of the distances of colors: intra-region visual error, and inter-region visual error.

The average consistency within each region of an image is calculated using the intra-region visual error measure. This is done by counting the number of pixels that have a color that stands out from the rest of the area in which they are located. Based on the Euclidean distance calculated in CIELAB color space, the perceptual differences between the two colors are determined.

On the other hand, the inter-region visual error measure evaluates the standard deviation of differences between neighboring regions in a segmentation. The number of boundary pixels in each region with an average color similar to a neighboring region, making them perceptually indistinguishable, is used to quantify this differentiation.

The final measure is obtained by combining the intra-region visual error and inter-region visual error through addition. This additive combination provides a comprehensive evaluation of the overall visual error in the segmentation process.

5.3. Silhouette Coefficient

Another metric that doesn't require a reference image to evaluate is that of the Silhouette Coefficient, proposed by Rousseeuw (1987). This metric was originally aimed at evaluating the similarity of a point to other points in its cluster, as opposed to every other point in different clusters, providing a way to assess clustering quality. In the context of image segmentation, each

pixel in the image would represent such a point assigned to a region. To define the overall coefficient, we first define the silhouette value for one single pixel in Equation 32.

$$s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))}, \quad (32)$$

where i is the current pixel being taken into consideration, $a(i)$ is the distance from that pixel to all other pixels belonging to the same region, and $b(i)$ is the average distance from that point to all other points belonging to the nearest neighboring region.

One thing to note is that if the value $s(i)$ is close to 1, then the pixel is considered to be assigned to the right region. However, if it is closer to 0, it is considered to be at a region boundary, whilst a negative value might indicate that the pixel was assigned to the wrong region.

Based on the aforementioned single-point silhouette formula, we can define the global Silhouette Coefficient as:

$$SC = \widetilde{s(k)}, \quad (33)$$

where $\widetilde{s(k)}$ represents the mean $s(i)$ across all pixels in a region, for a specific number of regions k.

The coefficient offers several upsides, in the sense that it is easy to interpret, with positive values indicating good segmentation, and negative values indicating bad segmentation. Similarly, it has very good general applicability as a metric since it doesn't require a ground truth reference in order to determine the quality of a segmentation, thus it can be applied across a variety of different algorithms.

However, it does suffer from several downsides, such as it being insensitive to the cluster shape, or the chosen number of regions. Due to the fact that the metric is based on minimising distances between pixels in the same regions and opposite regions, it assumes that the best segmentations are those made in the shape of circles, which is not common in image segmentation, as most regions tend to be irregularly shaped. Similarly, if the value k that is chosen for segmentation is not appropriate, some regions will display much narrower silhouettes when compared to the rest, thus choosing an appropriate value for k is of significant importance.

Additionally, since it requires pairwise computation for each point in order to give a final score, it tends to be computationally expensive, which made people implement simplified methods to compute faster such as Simplified Silhouette (Van der Laan et al., 2003) or Medoid Silhouette (Lenssen & Schubert, 2022). Last but not least the metric tends to perform poorly in the case of class imbalance, or if the number of classes is relatively small.

In recent years the coefficient is still used, but it tends to be used more in complement with other metrics, due to its limitation. It tends to be combined with metrics such as the Jaccard Index, Inertia, or the Davies-Bouldin Index, to avoid issues related to its sensitivity to cluster shapes, or computational cost for large datasets. This combination is also beneficial in the context of deep learning approaches, where several criteria such as cohesion, separation and overall similarities between the predicted regions and ground truth are of significant importance.

5.4. Inerția

Inertia, also known as Within-Cluster Sum of Squares (WCSS) is a metric that was originally introduced to evaluate the compactness of clusters (Macqueen 1967) and works similarly to the Silhouette Coefficient, with the main difference that it measures how tightly pixels are grouped in their region, based on the distance between each pixel and the region's centroids.

$$Inertia = \sum_{k} \sum_{x \in C_k} \|x - \mu_k\|^2, \quad (34)$$

where k represents the index of the current region being considered, C_k represents the pixels in region k and μ_k is the centroid of the region.

Lower inertia values tend to indicate that pixels in a region are more tightly grouped around their centroids and that there is a lower variance between the pixels in each region. However, one of the main limitations of inertia is that it doesn't consider distances between different regions, just the distances for pixels in the same region. Thus, it tends to increase the overall number of regions, as this would minimise the value, and lead to over-segmentation.

Similarly, like other metrics, it performs poorly for irregularly shaped regions, or when the image presents class imbalances. For these reasons, it is usually combined with other metrics, such as the Silhouette Coefficient or the Davies-Bouldin Index, to evaluate the quality of the segmentation, especially in applications where boundaries between regions are important.

5.5. Davies-Bouldin Index

Another metric that was proposed for evaluating the quality of segmented regions in an unsupervised manner is that of the Davies-Bouldin Index (Davies & Bouldin, 1979). It is used to measure the average similarity of each region when compared to its most similar region, considering both the compactness of the region and the separation between them. It is defined as:

$$DB = \frac{1}{K}\sum_{i=1}^{K} \frac{s_i + s_j}{d_{ij}}, \qquad (35)$$

where K represents the total number of regions, i and j are two non-overlapping regions, s represents the average distance between the pixels in a region, or in other words a measure of compactness, and d_{ij} represents the distance between the centroids of regions i and j, or a measure of separation.

Similar to inertia, the value of DB wants to be minimised, a lower value indicating a better separation of regions. For this reason, it tends to have the same advantages and disadvantages as some of the other aforementioned metrics.

Something to note is that the index considers non-overlapping regions. Whilst, usually, this could be considered a downside, as in practice clusters could present some overlaps, in the case of image segmentation, and especially unsupervised approaches, the regions are guaranteed to contain no overlap, thus this is an upside to the metric. Another advantage of the metric is because it only computes distances between pixels in a region, and only one distance between separate regions, it tends to have a low computation cost when compared to other metrics of similar nature.

Regarding outliers such as class imbalance, or irregular shapes, it also tends to favor regions of a more circular nature, thus it is very sensitive when the regions in the image diverge from this assumption. Similarly, a low score could also not be necessarily indicative of good segmentation. If we consider a greater distance between the centroids of each region, whilst maintaining shapes that favor lower average distances between the pixels of each region, the value DB will get closer to 0. This behavior would however lead to the under-segmentation of the image, thus picking an appropriate number of regions K that doesn't minimise it artificially is of important significance.

For these reasons, the Davies-Bouldin Index tends to be grouped with other metrics to properly balance the results, as some metrics, such as Inertia, favor over-segmentation, while others, such as DB, favor under-segmentation. Using this combination of metrics helps mitigate the weakness of any particular one, as this approach not only ensures that the segmentation minimises the distances between regions, but also maintains appropriate boundaries between them (Jain, 2010).

6. Conclusion

Despite the multitude of proposed image segmentation algorithms, consistently achieving human-level accuracy remains a major challenge. Advancement in this domain hinges on the

capability to verify if a new algorithm, or an alteration to a current one, truly signifies an improvement. This process is made easier by techniques for evaluating segmentation since they offer a comparison ground for different segmentation algorithms and their agreement with human perceptual classification.

Significant advantages can be attained via a segmentation evaluation process that is reliable. It might reveal traits that characterise a reliable segmentation method. It allows researchers to confirm claims that a newly proposed segmentation algorithm performs better than others. It can also assist application developers in selecting the most effective approach for their specific application.

Comparing a segmentation to a ground truth—an existing benchmark segmentation—is how supervised evaluation processes for segmentation work. The degree to which an image segmentation algorithm replicates human perceptual categorization can be determined through supervised evaluation once this ground truth has been manually determined. This strategy does, however, have certain limitations. When evaluating a segmentation algorithm, some room for improvement is suggested by the hierarchical pattern of perceptual grouping. Unfortunately, these accommodations frequently prevent us from determining the extent of under- or over-segmentation.

Direct comparison of the grouping hierarchies could be a more useful strategy for comparing automated segmentation with perceptual grouping. Several segmentation algorithms are capable of being changed to create a tree that shows different segmentation levels of detail. It would be interesting to compare this tree to one made similarly by an individual. However, neither a hierarchical dataset aimed at human perceptual categorization nor methods to evaluate the similarity of grouping trees have been developed.

Unlike supervised methods, unsupervised segmentation algorithms assess the quality of segmentation without relying on reference data, making them particularly useful in scenarios where ground truth is inconsistent or unavailable. While this represents a tougher task, there are various possible rewards. It avoids the aforementioned problem with inconsistent ground truth because it does not rely on a reference. Furthermore, it enables the unsupervised assessment to pick an acceptable parameter for a method automatically. The emphasis shifts between the workings and algorithmic details to the ultimate goal of segmentation - the ability to quantify segmentation quality without a guide, indicating a fundamental knowledge of human perceptual categorization.

A review article by Zhang et al. (2008) highlighted several key challenges related to unsupervised segmentation evaluation, denoting that many existing measurements at the time couldn’t distinguish between machine- and human-produced segmentations. While this provides an important foundation, significant advancements have been made in recent years, particularly in the context of integrating machine learning and deep learning techniques, which lead to more accurate and robust methods. Though segmentation evaluation is less mature than image segmentation, it has made significant strides in recent years. Several robust evaluation metrics now exist for both supervised and unsupervised segmentation. Additionally, publicly available datasets offer multiple ground truth segmentations for diverse images. Despite these advancements, more work remains to be done, particularly in developing and refining unsupervised metrics and developing high-quality datasets and metrics for interactive segment evaluation.

Acknowledgment

This work is part of the PhD thesis titled “Blending Technology and Education: A Deep Dive into Document Image Analysis, Image Processing, and Engineering Education” conducted by the first author.

References

Abraham, N., & Khan, N. M. (2019, April). A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th International Symposium on*

Biomedical Imaging (ISBI 2019) (pp. 683-687). IEEE. https://doi.org/10.1109/ISBI.2019.8759329

Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., & Blaschko, M. B. (2019). Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22* (pp. 92-100). Springer International Publishing.

Bharathi, S. (2019). A review on evaluative measures in image segmentation. *International Journal of Advance Research and Innovative Ideas in Education.*

Bowen, C., Ross, G., Piotr, D., Alexander, C. B., & Alexander, K. (2021). Boundary IoU: Improving object-centric image segmentation evaluation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15334-15342. https://doi.org/10.1109/CVPR46437.2021.01508

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6), 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

Cardoso, J. S., & Corte-Real, L. (2005). Toward a generic evaluation of image segmentation. *IEEE Transactions on Image Processing*, 14(11), 1773–1782. https://doi.org/10.1109/TIP.2005.854491

Chabrier, S., Emile, B., Rosenberger, C., & Laurent, H. (2006). Unsupervised performance evaluation of image segmentation. *EURASIP Journal on Advances in Signal Processing*, 2006(1). https://doi.org/10.1155/ASP/2006/96306

Chen, H.-C., & Wang, S.-J. (2004). The use of visible color difference in the quantitative evaluation of color image segmentation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 3, pp. iii–161). IEEE. https://doi.org/10.1109/ICASSP.2004.1326614

Cheng, B., Girshick, R., Dollár, P., Berg, A. C., & Kirillov, A. (2021). Boundary IoU: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15334-15342). https://doi.org/10.1109/CVPR46437.2021.01508

Correia, P. L., & Pereira, F. (2003). Objective evaluation of video segmentation quality. *IEEE Transactions on Image Processing*, 12(2), 186–200. https://doi.org/10.1109/TIP.2002.807355

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2), 224-227. https://doi.org/10.1109/TPAMI.1979.4766909

Hashemi, S. R., Salehi, S. S. M., Erdogmus, D., Prabhu, S. P., Warfield, S. K., & Gholipour, A. (2018). Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. *IEEE Access*, 7, 1721-1735. https://doi.org/10.1109/ACCESS.2018.2886371

Hatamizadeh, A., Terzopoulos, D., & Myronenko, A. (2019). End-to-end boundary aware networks for medical image segmentation. In *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10* (pp. 187-194). Springer International Publishing. https://doi.org/10.1007/978-3-030-32692-0_22

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *arXiv preprint arXiv:1703.06870.* https://doi.org/10.48550/arXiv.1703.06870

Huang, Q., & Dom, B. (2002). Quantitative methods of evaluating image segmentation. *Proceedings of the International Conference on Image Processing*, Washington, DC, USA. https://doi.org/10.1109/ICIP.1995.537578

Itaya, Y., Tamura, J., Hayashi, K., & Yamamoto, K. (2024). Asymptotic properties of Matthews correlation coefficient. *arXiv preprint arXiv:2405.12622*. https://doi.org/10.48550/arXiv.2405.12622

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

Jiang, X., Marti, C., Irniger, C., & Bunke, H. (2006). Distance measures for image segmentation evaluation. *EURASIP Journal on Advances in Signal Processing, 2006*(1), Article 35909. https://doi.org/10.1155/ASP/2006/35909

Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019). Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9404-9413). https://doi.org/10.1109/CVPR.2019.00963

Lenssen, L., & Schubert, E. (2022, September). Clustering by direct optimization of the medoid silhouette. In *International Conference on Similarity Search and Applications* (pp. 190-204). Cham: Springer International Publishing. https://doi.org/10.48550/arXiv.2209.12553

Li, Y., Xue, Z., & Xu, H. (2024). OTAS: Unsupervised boundary detection for object-centric temporal action segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 6437-6446). https://doi.org/10.48550/arXiv.2309.06276

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, 13 (pp. 740-755). Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48

Lin, Y., Zhang, D., Fang, X., Chen, Y., Cheng, K. T., & Chen, H. (2023, June). Rethinking boundary detection in deep learning models for medical image segmentation. In *International Conference on Information Processing in Medical Imaging* (pp. 730-742). Cham: Springer Nature Switzerland. https://doi.org/10.48550/arXiv.2305.00678

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2002). A database of human-segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. Presented at the Eighth IEEE International Conference on Computer Vision, Vancouver, BC, Canada. https://doi.org/10.1109/iccv.2001.937655

McGuinness, K. (2009). *Image Segmentation, Evaluation, and Applications*. School of Electronic Engineering, DCU.

Monteiro, F. C., & Campilho, A. C. (2012). Distance measures for image segmentation evaluation. Presented at the *Numerical Analysis and Applied Mathematics ICNAAM 2012: International Conference of Numerical Analysis and Applied Mathematics, Kos, Greece*. https://doi.org/10.1063/1.4756257

Ntirogiannis, K., Gatos, B., & Pratikakis, I. (2013). Performance evaluation methodology for historical document image binarization. *IEEE Transactions on Image Processing*, 22(2), 595–609. https://doi.org/10.1109/TIP.2012.2219550

Ogwok, D., & Ehlers, E. M. (2022). Jaccard Index in Ensemble Image Segmentation: An Approach. *Proceedings of the 2022 5th International Conference on Computational Intelligence and Intelligent Systems*. Presented at the CIIS 2022: The 5th International Conference on Computational Intelligence and Intelligent Systems, Quzhou, China. https://doi.org/10.1145/3581792.3581794

Pinastawa, I. W. R., Pradana, M. G., & Khoironi, K. (2024). Edge detection model performance using Canny, Prewitt, and Sobel in face detection. *Sinkron*, 8(2), 623–631. https://doi.org/10.33395/sinkron.v8i2.13497

Prewitt, J. M. S., & Mendelsohn, M. L. (1966). The analysis of cell images. *Annals of the New York Academy of Sciences*, 128(3), 1035–1053.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Lecture Notes in Computer Science* (pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 410–420).

Rother, C., Kolmogorov, V., & Blake, A. (2004). 'GrabCut': Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Transactions on Graphics*, 23(3), 309–314. https://doi.org/10.1145/1015706.1015720

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In *Wang, Q., Shi, Y., Suk, H. I., & Suzuki, K. (Eds.) Machine Learning in Medical Imaging. MLMI 2017. Lecture Notes in Computer Science*, vol. 10541. Springer, Cham. https://doi.org/10.1007/978-3-319-67389-9_44

Setiawan, A. W. (2020). Image segmentation metrics in skin lesion: Accuracy, sensitivity, specificity, dice coefficient, Jaccard index, and Matthews correlation coefficient. *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*. Presented at the 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), Surabaya. https://doi.org/10.1109/cenim51130.2020.9297970

Sun, F., Luo, Z., & Li, S. (2023). Boundary Difference over Union Loss for Medical Image Segmentation. In *Lecture Notes in Computer Science* (pp. 292–301). https://doi.org/10.48550/arXiv.2308.00220

Van der Laan, M., Pollard, K., & Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, *73*(8), 575–584. https://doi.org/10.1080/0094965031000136012

Wang, X. (2007). Laplacian operator-based edge detectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5), 886–890. https://doi.org/10.1109/TPAMI.2007.1027

Xia, F., Liu, J., Nie, H., Fu, Y., Wan, L., & Kong, X. (2020). Random walks: A review of algorithms and applications. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(2), 95–107. https://doi.org/10.1109/tetci.2019.2952908

Yang, L., Albregtsen, F., Lønnestad, T., & Grøttum, P. (1995). A supervised approach to the evaluation of image segmentation methods. In *Computer Analysis of Images and Patterns* (pp. 759–765). https://doi.org/10.1007/3-540-60268-2_377

Yang, H. (2023). Graphic image segmentation method based on high-precision and fast algorithm. *2023 Global Conference on Information Technologies and Communications (GCITC)*, 40, 1–4. Presented at the 2023 Global Conference on Information Technologies and Communications (GCITC), Bangalore, India. https://doi.org/10.1109/gcitc60406.2023.10425884

Yap, P. T., & Raveendran, P. (2004). Image focus measure based on Chebyshev moments. *IEE Proceedings - Vision, Image and Signal Processing*, 151(2), 128.

Zhang, H., Cholleti, S., Goldman, S. A., & Fritts, J. E. (2006). Meta-evaluation of image segmentation using machine learning. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*. Presented at the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06), New York, NY, USA. https://doi.org/10.1109/cvpr.2006.185

Zhang, H., Fritts, J. E., & Goldman, S. A. (2008). Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110(2), 260–280. https://doi.org/10.1016/j.cviu.2007.08.003

Zhang, H., Fritts, J. E., & Goldman, S. A. (2005). A co-evaluation framework for improving segmentation evaluation. In I. Kadar (Ed.), *Signal Processing, Sensor Fusion, and Target Recognition XIV*. https://doi.org/10.1117/12.604213

Zhang, Y. J. (1996). A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8), 1335–1346. https://doi.org/10.1016/0031-3203(95)00169-7

Zhao, H., Mao, R., Li, M., Li, B., & Wang, M. (2023). SheepInst: A high-performance instance segmentation of sheep images based on deep learning. *Animals*, 13(8), 1338.