


Article

# Assessment of the Segmentation of RGB Remote Sensing Images: A Subjective Approach

Giruta Kazakeviciute-Januskeviciene <sup>1,2,\*</sup>, Edgaras Janusonis <sup>1</sup>, Romualdas Bausys <sup>1</sup> , Tadas Limba <sup>2</sup> and Mindaugas Kiskis <sup>2</sup>

<sup>1</sup> Department of Graphical Systems, Vilnius Gediminas Technical University, Sauletekio al. 11, LT-10223 Vilnius, Lithuania; edgaras.janusonis@stud.vgtu.lt (E.J.); romualdas.bausys@vgtu.lt (R.B.)

<sup>2</sup> Faculty of Public Governance and Business, Mykolas Romeris University, Ateities str. 20, LT-08303 Vilnius, Lithuania; tlimba@mruni.eu (T.L.); mkiskis@mruni.eu (M.K.)

\* Correspondence: giruta.kazakeviciute-januskeviciene@vgtu.lt

Received: 31 October 2020; Accepted: 15 December 2020; Published: 18 December 2020



**Abstract:** The evaluation of remote sensing imagery segmentation results plays an important role in the further image analysis and decision-making. The search for the optimal segmentation method for a particular data set and the suitability of segmentation results for the use in satellite image classification are examples where the proper image segmentation quality assessment can affect the quality of the final result. There is no extensive research related to the assessment of the segmentation effectiveness of the images. The designed objective quality assessment metrics that can be used to assess the quality of the obtained segmentation results usually take into account the subjective features of the human visual system (HVS). A novel approach is used in the article to estimate the effectiveness of satellite image segmentation by relating and determining the correlation between subjective and objective segmentation quality metrics. Pearson's and Spearman's correlation was used for satellite images after applying a k-means++ clustering algorithm based on colour information. Simultaneously, the dataset of the satellite images with ground truth (GT) based on the "DeepGlobe Land Cover Classification Challenge" dataset was constructed for testing three classes of quality metrics for satellite image segmentation.

**Keywords:** satellite image segmentation; segmentation quality assessment; correlation analysis; objective quality metrics; subjective evaluation

## 1. Introduction

Satellite imagery classification is one of the main tasks in remote sensing applications such as city planning, climate change research [1], earth observations, geographical maps improvement [2], topographic surveys, or for the military purposes. Colour segmentation may also be used to track land cover changes over time [3]. The most prominent goal of remote sensing data analysis is object detection, based on image segmentation [2]. The principal goal of the image segmentation process is to partition the image into a set of segments that are homogeneous, according to specific criteria such as colour spectrum, shape, intensity, or texture [3] and map the individual regions to the corresponding real-world objects, like rivers, fields, roads, and other. Image segmentation is a broad term dependent on the goals of the specific application. Image segmentation is utilized for a wide variety of applications and is also a part of the object detection process.

Remote sensing images (RSI) usually represent various parts of the Earth surface and are characterized by clear details along with diverse spatial and textural information. When compared to most natural images, the satellite images stand out as being highly structured and uniform [1].

While improving current state-of-the-art segmentation methods, it is equally important to have proper methods for segmentation quality evaluation. The performance of colour segmentation greatly impacts the quality of an image understanding system [4].

Since the subjective quality assessment is expensive, resource-intensive, and time-consuming, the aim is to investigate the correlation between objective quality assessment methods (QAM) and the human visual system (HVS) so that appropriate objective QAM can be applied to assess the quality of satellite image segmentation.

A wide variety of objective metrics for examining segmentation quality are proposed. Fewer studies were conducted to compare the correlation of widely used objective segmentation metrics to the subjective quality assessment. Usually, only a small subset of available quality metrics is being compared against new quality measures during their development process [5–7].

The authors proposed to relate various objective metrics to the subjective scores by using metric classes, which is a novel approach in the image segmentation. The defined metrics classes were tested with the constructed dataset of the satellite images with ground truth (GT). The DeepGlobe Land Cover dataset was used as the base for the construction of our dataset. The original results of this study can be applied for the development of image segmentation quality analysis or combined with the related knowledge for the derived results, such as selecting a quality metric for a particular dataset or application that demands both tight correlations with human perception and low computational complexity. Results can also help suggest a possible combination of different metrics for achieving a more accurate relation with human perception of segmented images.

This article has the following structure. Section 2 provides a summary on published papers in the context of subjective and/or objective segmentation quality assessment of the satellite or natural images. Section 3 provides an overview of the k-means++ clustering method. Section 4 describes the characteristics of dataset and dataset preparation for the segmentation, objective, and subjective quality assessment. Internal, external, and full-reference image quality assessment (FR-IQA) metric groups considered in this experiment are explained in Section 5. Section 6 provides information on the subjective quality assessment procedure. Experimental results, including the roadmap of technology and used correlation coefficients, are presented in Section 7. Discussion and directions for future research are presented in Section 8. Finally, conclusions in Section 9 summarize the main observations from the discussion.

## 2. Related Works

Over many years, the development of image quality assessment (IQA) has drawn extensive and constant attention. Relatively less research was conducted from the perspective of evaluating image segmentation quality, especially the way humans understand and estimate the quality of segmented images.

Authors of Reference [7] systematized human visual properties, important for designing an objective segmentation metric: (1) human visual tolerance (e.g., for border distortions), (2) human visual saturation (difficulty for the human to evaluate similarity when distortions become large), (3) different perception of false negative (FN) and false positive (FP) pixels, and (4) the strongest distortions determine overall segmentation quality.

For natural images, the authors of Reference [8] suggested a contour-based score, combining Jaccard index (*JI*) and *BF* (Boundary *F1*) score for increased correlation with human rankings, as *JI* does not take into account quality of segmentation boundaries. It was observed that, while the accurate contours are definitely less important than correct classification, the proposed measure can further improve correlation with human rankings for high-quality segmentation. Authors used PASCAL VOC 2007 and 2011 datasets. However, very imbalanced data in PASCAL Visual Object Classes (PASCAL VOC) dataset can create biased results for some metrics like Accuracy (*ACC*). Ground truths contain instances of individual objects labelled separately in addition to the auxiliary label for object contours.

Similar to Reference [8], authors in Reference [9] proposed a subjective quality metric for single object segmentation combining region and boundary-based terms that relate to human visual tolerance and saturation properties. A subjective test showed the improved results, achieving a Spearman Rank Order Correlation Coefficient (*SROCC*) value of 0.88 and then compared to *Jl*, *F1* score (*F1*), Fuzzy Contour (*FC*), *BF* score, and Mixed Measure (*MM*) metrics. For testing the new metric, the dataset was created by selecting images from other object segmentation databases like Microsoft research Asia salient object database, PASCAL VOC 2012, and Microsoft research Cambridges grabcut database.

The effectiveness of Peak signal-to-noise ratio (*PSNR*) as a quality measure for image segmentation algorithms was explored in Reference [10]. In this experiment, GT data was created by modifying images from the Berkeley BSR300 database (intended for evaluating edge detection algorithms) to resemble threshold segmentation. *PSNR* was used to measure the similarity between the GT image and segmentation result obtained by applying salt and pepper noise on the created GT image. It was noted that quality of the edge detection can be more effectively evaluated by *PSNR*. The tests performed by the authors, however, have not considered multi-region segmentation, which is a more practical approach for real-world objects.

Evaluation of segmentation quality for satellite images is regarded as a difficult task since there is no universal standard for evaluating segmentation results of satellite images. The most common evaluation methods are based on external or full-reference quality measures, and the need of sufficient amount of reference data poses a problem. Authors of Reference [11] suggested a Synthetic Image Testing Framework (SITEF) for the evaluation of multispectral satellite image segmentation by using synthetic images. This method provides different evaluation perspectives such as parcel size, shape, and land cover type. The framework was tested using images obtained by SPOT HRG satellite consisting of six land cover classes. The Hammoude metric, Rand coefficient, Corrected Rand coefficient, and *Jl* were used for segmentation evaluation.

Authors in Reference [12] proposed attention dilation-LinkNet (AD-LinkNet) neural network that displays a significant improvement on the segmentation accuracy of satellite images. Three satellite segmentation datasets were used: DeepGlobe's road extraction and land cover datasets [1] as well as Inner Mongolia's land classification dataset. For different data sets, different model optimizations are needed. *Jl* was used for evaluating both road and land segmentation results.

There are apparent differences between satellite images [1] and natural images, as seen in PASCAL VOC 2012 dataset. In satellite imagery, every object has a semantic meaning, while natural images are more chaotic than satellite, and often include large areas of background, which is less important when compared to foreground objects.

Colour image segmentation can provide more information for detecting objects than using grayscale images. Satellite images consist of homogenous colour regions, that define image data and grouping of pixels can be performed on distinct colour features. Many papers have been published in the past, focusing on using colour features for segmenting satellite images [3,13]. Automatic detection of road segments from RSI for road detection applications [14] suggested an approach that can also be applied to detect other objects in RSI. Authors in Reference [15] proposed a method for land cover classification based on colour moments (mean, standard deviation, and skewness) used in extracting colour features. The segmentation method in Reference [16] used wavelet transform for extracting colour and texture features from satellite images in the YCbCr colour space.

### 3. Image Segmentation

A wide variety of clustering solutions have been proposed for solving various problems depending on application-specific goals and/or characteristics of a particular dataset. A cluster is a group of pixels, which are similar and also dissimilar to pixels in other clusters, according to specific image features, that can be used for further image analysis. If the segmentation uses colour as a feature, then pixels are assigned to the clusters based on their colour similarity. Additionally, segmentation can be performed

by using a combination of different features [16]. Using segmentation by colour, we can also avoid feature calculation for every pixel, thus, improving overall performance.

For this research, the k-means [17] clustering algorithm was selected. Researchers tend to employ clustering methods with well-known advantages and disadvantages, avoiding undiscovered limitations that could impact results [18]. K-means popularity and its simple implementation make it easier for replicating the same experiment independently. However, other clustering algorithms can be used, which might provide a different perspective and/or better results in certain situations (e.g., finding more complex cluster shapes).

The k-means algorithm (Table 1) groups the  $N$  data points or pixels  $\{x_1, x_2, \dots, x_i, \dots, x_N\}$  into  $K$  clusters by minimizing its objective function  $SSE$  (sum of squared error) within a cluster for all clusters. Minimizing  $SSE$  also maximizes  $SSB$  (the sum of squares between clusters). For this reason,  $SSE$ ,  $SSB$ , and metrics combining these criteria return higher scores for clusters constructed by k-means.

**Table 1.** The k-means clustering method.

Step	Description
1	An optimal number of clusters $K$ is selected by the Silhouette method [19].
2	Instead of random initialization, cluster centroids are initialized using k-means++ procedure (Table 2).
3	The Euclidean distance $D = d(x_i, c_i)$ is calculated between the cluster centroids and each pixel of an image.
4	Based on the calculated $D$ , all pixels are assigned to the nearest centroid.
5	Recalculate all cluster centroid positions $c_i$ by computing the mean of currently assigned pixels. $c_i = 1/ C_i  \cdot \sum_{x \in C_i} x_i$
6	The cycle is repeated (from the third step) until the position of the cluster centroids no longer changes <sup>1</sup> (i.e., no pixels are reassigned).

<sup>1</sup> Alternatively, another criterion, e.g., (1) a limited number of iterations, (2) insignificant changes of cluster centroids positions, and (3)  $SSE$  falls below the set limit—possible combinations.

**Table 2.** The k-means++ cluster centroid initialization method [20].

Step	Description
1	First cluster centroid $c_1$ is selected evenly randomly from the existing set of pixels $X$ .
2	Calculate distances $D(x)$ from each pixel to the nearest centroid (in the first iteration, this is $c_1$ ).
3	Each subsequent cluster centroid $c_i$ is selected from the remaining set of pixels $x \in X$ with probability: $\frac{D^2(x)}{\sum_{x \in X} D^2(x)}$
4	Go back to step 2 and repeat $K-1$ times until $K$ cluster centroids have been added.
5	Proceed with step 3 from Table 1.

Before the segmentation stage, no preprocessing is necessary, although, according to Reference [8], the quality metric scores might have less meaning when the quality of segmentation is too low, suggesting that middle-quality segmentation is the best choice for the subjective quality evaluation. As such, we selected the optimal number of clusters by the Silhouette method [19] using a k-means++ algorithm to find their initial centroids (Table 2).

Results produced by k-means are dependent on the initialization procedure of cluster centroids. It is important to configure parameters for more deterministic behavior, not to impact the calculations of the metrics severely. The prepared images from our satellite dataset (described in Section 4) have lower resolution ( $324 \times 220$  px) and fewer clusters (2–4). We observed segmentation results returned by MATLAB implementation of k-means seem to be stable even with the default value of 100 iterations and only three replications. Alternatively, it is possible to (1) lock cluster centroids by selecting them manually or (2) use Pseudo-Random Number Generator (*PRNG*) with constant seed [18]. In our case, k-means++ initialization (Table 2) was used, which also can improve segmentation results.

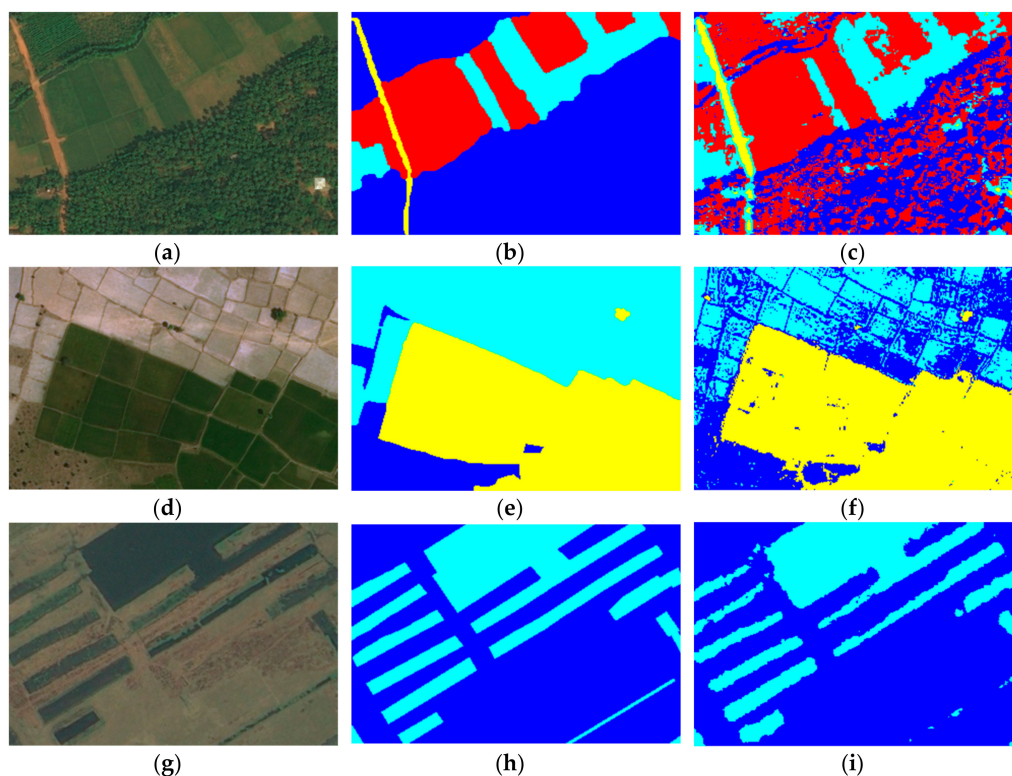
Satellite images are segmented in the CIE 1976 L\*a\*b\* colour space, as suggested in Reference [14], which also more accurately represents the human perception of the colour [21] and have better performance than RGB space in many colour image applications. The Euclidean distance is used to measure colour similarities between pixels in the a\*b\* plane [3].

#### 4. Satellite Images Dataset

There are many datasets designed for image segmentation problems. The selection of an appropriate dataset is a crucial decision that impacts subsequent choices, such as the selection of the optimal clustering method. It is important to understand the limitations and possible problems of the particular dataset before using it for any research project.

The majority of quality metrics for the assessment of the segmented image quality requires the GT image. For the evaluation of image quality after such distortions as image compression, the GT is treated as the original image, and the GT preparation process is unnecessary. Obtaining GT is always a barrier for automated image segmentation. Human skills are often required for manual labelling, and it is a time-consuming process. The manual labelling itself is subjective, and the different GT versions of the same image may be produced. Widely used datasets like Berkeley segmentation datasets (BSDS) [22] often contain natural images very diverse in their content, but they do not necessarily serve as a target for the specific application or provide GT, which match specific segmentation goals.

Figure 1 depicts images from our constructed dataset based on the “DeepGlobe Land Cover Classification Challenge” dataset [1] intended to solve a mentioned problem by providing satellite imagery with GT data for improving state-of-the-art satellite image processing methods.



**Figure 1.** Segmentation results for satellite images from our dataset part 1 [23] (different clusters for each image are marked with selected colours): (a) satellite image “71619\_sat,” (b) corrected GT image “71619\_gt” with 4 clusters, (c) k-means++ segmentation results “71619\_seg” of “71619\_sat,” (d) satellite image “161109\_sat,” (e) corrected GT image “161109\_gt” with three clusters, (f) k-means++ segmentation results “161109\_seg” of “161109\_sat,” (g) satellite image “676758\_sat,” (h) corrected GT image “676758\_gt” with two clusters, (i) k-means++ segmentation results “676758\_seg” of “676758\_sat.”.

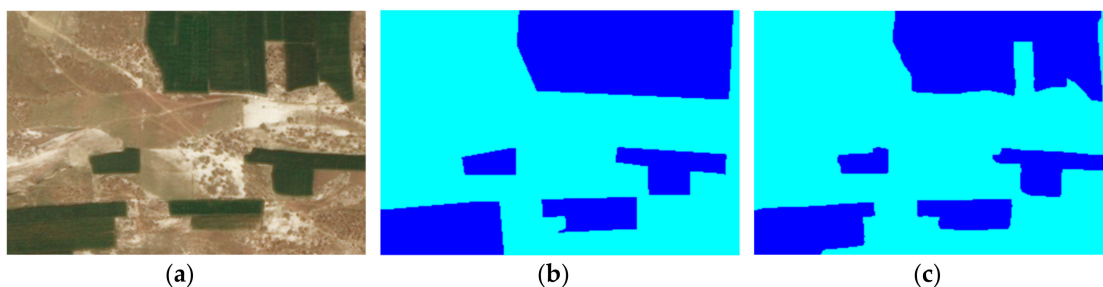
The original dataset is divided into three parts: test dataset (172 images), validation dataset (171 images), and training dataset, which consists of satellite images paired with corresponding GT images. Training dataset includes 1606 images in total collected by DigitalGlobe's satellite: 803 satellite images in RGB, a size of  $2448 \times 2448$  px 24-bit JPEG image format and 803 GT images in RGB, and a size of  $2448 \times 2448$  px 24-bit PNG image format. Files are named using the following format: <ID>\_sat.jpg for satellite images and <ID>\_mask.png for GT images. GT images contain seven land cover types: (1) urban land, (2) agriculture land, (3) forest land, (4) water, (5) barren land, and (6) rangeland and unknown (e.g., clouds). Each cover type is described in (R, G, B) colour codes. Dataset was sampled for land cover classes to have enough representation with agriculture class having 56.76% of total pixel count and water class having at least 3.74%.

The computation time of quality metrics depends on the size of input images. High computational complexity metrics like Silhouette [24] are expensive when calculating distances. For this reason,  $324 \times 220$  px regions and corresponding GT regions were cropped from the existing  $2448 \times 2448$  px images and corresponding GT images, provided by a DeepGlobe Land Cover dataset (Figure 2). This also limits the number of possible clusters, making the land cover objects evaluated less distracting for human observers.



**Figure 2.** Sampled satellite images from the DeepGlobe Land Cover dataset with two selected  $324 \times 220$  regions that are highlighted in red and are used for our dataset: (a) “676758\_sat.jpg,” (b) “762359\_sat.jpg,” (c) “941237\_sat.jpg,” (d) “668465\_sat.jpg.” Names of images from our constructed dataset [23] share the same <ID's> in part1 and part2 folders.

The cost for annotating multi-class segmentation GT images provided in the DigitalGlobe Land Cover dataset have segments that are less accurate, missing prominent land cover portions (Figure 3). Therefore, semi-automatic adjustments to the GT images were performed inside MATLAB using an Image Labeler, which allows us to label image data manually or use automation, and to export to the MATLAB workspace as a ground truth object variable containing label definitions.



**Figure 3.** Example of visible structural inaccuracies in agriculture land (different image clusters are marked with selected colours): (a) satellite image “965977\_sat\_00000.png” from our constructed dataset, (b) not corrected GT image of “965977\_sat\_00000.png,” and (c) corrected GT image of “965977\_sat\_00000.png” from our constructed dataset.

The constructed dataset, including average mean opinion score (MOS) scores for each image used in our survey, can be accessed at Reference [23].

## 5. Objective Quality Assessment Methods for Image Segmentation

Depending on whether segmentation results are evaluated by a human or an algorithm, quality assessment is divided into two main branches: objective and subjective [25,26]. The main intention of objective quality models is to approximate properties of HVS in order to avoid slow and impractical subjective testing procedures. For this reason, the design process of new objective quality metrics often includes correlation tests with an obtained MOS [5,6].

The objective metrics may require initial information to evaluate the image segmentation quality. Such information is known as a reference image or GT. It could be prepared manually so that a comparison could be made with segmentation results achieved by a particular algorithm. This group of metrics is called external metrics (or supervised evaluation measures). The external information may not always be available, and metrics that do not depend on external information are used. This group of metrics is called internal metrics (or unsupervised evaluation measures).

### 5.1. External Metrics

The evaluation of segmented image quality using external metrics is equivalent to comparing two segmentation versions (GT image and segmentation result), where each pixel has a unique class label (or index) assigned to it. The GT image is often created (labelled) by an expert in a particular field (e.g., medical image segmentation) or sometimes can be generated [11] from an input image in the form of synthetic information. In our case, the segmentation result is obtained from the clustering algorithm, which returns an array containing cluster indices of each pixel corresponding to which cluster that pixel was assigned. Then the quality of the segmentation result can be evaluated by an external metric taking GT image and segmentation result as an input to determine the level to which two segmentations match. The evaluation process of the external metric is based on the analysis of cluster indices assigned to all pixel pairs between the GT image and segmentation result.

External quality metrics are calculated from the confusion matrix. The confusion matrix is a summary of the results of the segmentation problem. The number of correct and incorrect assignments is summed for a specific class. The confusion matrix is defined as a square matrix ( $K \times K$ ) where  $K$  is a number of classes (or clusters) and consists of four parameters (Table 3). These parameters are then used to derive combined external metrics that are presented in Table 4.

It is worth noting that external metrics can also serve as a mean of comparing segmentation results of two different algorithms or segmentation results of a single algorithm, but with different parameters [24].

**Table 3.** Notation for the external comparison metrics.

Notation	Name	Description
<i>TP</i>	true positive pixels	pixels that belong to cluster $C_i$ in the GT image and are correctly assigned to cluster $C_i$ in the segmented image (i.e., common pixels between GT image and segmented image)
<i>TN</i>	true negative pixels	pixels that are assigned to different clusters both in the segmented and GT images
<i>FP</i>	false positive pixels	pixels that are incorrectly assigned to cluster $C_i$ in the segmented image compared to the GT image
<i>FN</i>	false negative pixels	pixels that are assigned to cluster $C_i$ in the GT image, but assigned to different cluster in the segmented image

Table 4. List of external metrics.

Symbols/Notations Commonly Used	Names Commonly Used	Definition	
ACC	Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	(1)
PPV, P	Positive Prediction Value, Precision	$\frac{TP}{TP+FP}$	(2)
TPR, R	True Positive Rate, Recall, Sensitivity	$\frac{TP}{TP+FN}$	(3)
TNR	Specificity, True Negative Rate	$\frac{TN}{TN+FP}$	(4)
MCC	Matthews correlation coefficient	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	(5)
JI, IoU	Jaccard index, Intersection over Union	$\frac{TP}{TP+FP+FN}$	(6)
$F_1$ , DSC, Dice	Sørensen–Dice coefficient	$2 \cdot \frac{P \cdot R}{P+R} = \frac{2TP}{2TP+FP+TP}$	(7)
$F_2$	-	$5 \cdot \frac{P \cdot R}{4P+R}$	(8)
$F_{1/2}$	-	$5 \cdot \frac{P \cdot R}{P+4R}$	(9)
KI	Kulczynski index [27]	$\frac{1}{2}(P+R) = \frac{1}{2}\left(\frac{TP}{TP+FP} + \frac{TP}{TP+FN}\right)$	(10)
FMI	Folkes–Mallows index [27]	$\sqrt{P \cdot R} = \frac{TP}{\sqrt{(TP+FP) \cdot (TP+FN)}}$	(11)

The  $F_\beta$  measure is defined as a weighted harmonic mean of  $P$  and  $R$ :

$$F_\beta = (1 + \beta^2) \frac{P \cdot R}{(\beta^2 \cdot P) + R} \quad (12)$$

In Equation (12), the parameter  $\beta$  is any real positive number ( $0 \leq \beta \leq +\infty$ ) and determines the weighting between  $P$  and  $R$ . If  $\beta > 1$ , higher weight is applied to  $R$ . If  $\beta < 1$ , higher weight is applied to  $P$ . Depending on the value of  $\beta$ , expression (12) can lead to several possible scenarios/metrics:

- if  $\beta \rightarrow +\infty$ ,  $\Rightarrow F_\beta = R$ ,
- if  $\beta = 2$ ,  $\Rightarrow F_\beta$  is equal to  $F_2$ , where  $R$  has double weight compared to  $P$ ,
- if  $\beta = 1$ ,  $\Rightarrow F_\beta$  is equal to the unweighted harmonic mean of  $P$  and  $R$ , and is equal to  $F_1$ . In this case,  $P$  and  $R$  are equally important,
- if  $\beta = 0.5$ ,  $\Rightarrow F_\beta$  is equal to  $F_{1/2}$  where  $R$  has half weight compared to  $P$ ,
- if  $\beta \rightarrow 0$ ,  $\Rightarrow F_\beta = P$ .

KI is defined as the arithmetic mean of  $P$  and  $R$ , while the FMI is defined as the geometric mean of  $P$  and  $R$ . Since the geometric mean is always in-between of arithmetic mean and harmonic mean for any positive number (in this case  $0 \leq P, R \leq 1$ ), then it is also true that  $K \geq FMI \geq F_1$ .

All external metrics listed in Table 4 have a range of [0; 1], except MCC, ranging [−1; 1], where 1 represents a perfect segmentation result. However, various research studies suggest [8,9] that region-based metrics like  $F_1$  or  $JI$  alone cannot fully reflect the human perception of image segmentation quality.

The definitions of metrics in Table 4 can only be applied to the binary (two-class) segmentation case. For multiclass segmentation, the overall scores for external measures were obtained by finding the score for each cluster  $C_i$  (function *confusionmatStats* [28]), and then calculating the unweighted mean among all clusters  $K$ .

$$overall\_external\_measure = \frac{1}{K} \sum_{C=1}^{C_K} external\_measure_C \quad (13)$$



## 5.2. Internal Metrics

Internal metrics require no external information, i.e., reference image, to evaluate the segmentation quality. The segmented result is evaluated based on a particular set of characteristics (criteria), derived from the initial dataset. This feature is important, as generating or creating reference images is time-consuming or sometimes impossible.

The internal quality metrics are usually employed for (1) solving an optimal number of clusters, (2) determining the quality of clustering results without depending on external information, and (3) determining if data have any structure [24].

Internal methods evaluate clustering by examining the separation and the compactness of the clusters.

- Cluster cohesion (or compactness) measures how closely related objects are in a cluster or how close the data points are from the cluster centroid. Better clustering results have pixel values close to their respective cluster centroids.
- Cluster separation measures how a cluster differs or is separated from the other clusters. Better clustering results have centroids of different clusters far from each other.

The primary measures of cohesion (14) and separation (15) are calculated from the image under investigation, while (16), (17), (19), and (21) measures combine both cohesion and separation. Basic notation for internal metrics is provided in Table 5. Clustering quality is considered good when the clusters are well separated and compact.

**Table 5.** Basic notations and definitions for the internal metrics.

Notation	Description
$K$	the number of clusters
$N$	the number of objects (pixels) in image $X$ (i.e., pixel count or resolution)
$\{C_1, C_2, \dots, C_i, \dots, C_K\}$	the set of $K$ clusters
$D = d(\cdot, \cdot)$	the Euclidean distance between two objects (pixels)
$ C_i $	the total number of data points (pixels) in a cluster $i$
$X_i = \{x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{i C_i }\}$	the set of pixels in $C_i$
$\{c_1, c_2, \dots, c_i, \dots, c_K\}$	the set of cluster means (centroids)

The Sum of Squared Errors Within Cluster ( $SSW$ ) is alternatively known as the Sum of Squared Errors ( $SSE$ ) (14) [29]. Lower values indicate higher cluster cohesion.  $SSE$  decreases with the increase of the number of clusters.  $SSE$  is defined as:

$$SSE_K = \sum_{i=1}^{i=K} \sum_{j=1}^{j=|C_i|} d(x_{ij}, c_i)^2 \quad (14)$$

The Sum of Squares Between Clusters ( $SSB$ ) (15) [29] is a measure of separation. Higher  $SSB$  values indicate more separated clusters.  $SSB$  is defined as the sum of squared distances from  $c$  (known as overall centroid, i.e., centroid of all cluster centroids) to other cluster centroids  $c_i$  each time multiplied by the number of pixels in the cluster  $C_i$ .

$$SSB_K = \sum_{i=1}^{i=K} |C_i| \cdot d(c_i, c)^2 \quad (15)$$

Using  $SSE$  and  $SSB$ , some other combined internal metrics can be calculated. The Calinski–Harabasz index ( $CHI$ ) (16) [30] is alternatively known as the variance ratio criterion ( $VRC$ ).

$$CH = \frac{SSB_K}{SSE_K} \cdot \frac{N - K}{K - 1} \quad (16)$$

The larger the  $SSB_K/SSE_K$  ratio is, the better the clustering quality is.

The Hartigan index ( $HI$ ) (17) is defined as the logarithmic relationship between  $SSB$  and  $SSE$  [31].

$$HI = \log\left(\frac{SSB_K}{SSE_K}\right) \quad (17)$$

Numbers in the range [0; 1] have negative logarithms. Therefore, if  $SSB > SSE$ , then  $HI > 0$ . If  $SSE > SSB$ , then  $HI < 0$ .  $HI = 0$  only if  $SSB$  is equal to  $SSE$ .

The  $X_u$  coefficient (18) combines  $SSE_K$ , number of clusters  $K$ , a total number of pixels, and dimensionality of input data [32].

$$X_u = D \cdot \log_2\left(\sqrt{\frac{SSE_K}{DN^2}}\right) + \log K \quad (18)$$

The Silhouette coefficient ( $SH$ ) (19) [19] is another popular way of combining cohesion and separation [27].

$$S(x_{ij}) = \frac{b(x_{ij}) - a(x_{ij})}{\max\{b(x_{ij}), a(x_{ij})\}} \text{ or } S(x_{ij}) = \begin{cases} 1 - \frac{a(x_{ij})}{b(x_{ij})}, & \text{if } a(x_{ij}) < b(x_{ij}), \\ 0, & \text{if } a(x_{ij}) = b(x_{ij}), \\ \frac{b(x_{ij})}{a(x_{ij})} - 1 & \text{if } a(x_{ij}) > b(x_{ij}). \end{cases} \quad (19)$$

The coefficient ( $-1 \geq S(x_{ij}) \geq 1$ ) can be calculated for an individual pixel  $x_{ij}$ .

- $a(x_{ij})$  average distance from the pixel  $x_{ij}$  to other pixels in  $C_i$  (cohesion),
- $b(x_{ij})$  is the minimum (worst case) of the all-average distances, each of them computed among the same pixel  $x_{ij}$  and all the pixels inside another cluster (separation).

$SH$  value for an individual pixel  $x_{ij}$  represents how similar  $x_{ij}$  is to pixels inside its own cluster, compared to pixels in other clusters. The Silhouette coefficient for a single cluster  $C_i$ .

$$S(C_i) = \frac{\sum_{j=1}^{|C_i|} S(x_{ij})}{|C_i|} \quad (20)$$

Finally, the overall  $SH$  for an image can be calculated similarly to Equation (13). However, alternative ways of calculating the overall score are possible (e.g., by averaging  $SH$  values for all pixels). Higher  $SH$  values indicate a better clustering result. More detailed interpretations of  $SH$  values are described in Reference [19].

The Davies-Bouldin index ( $DBI$ ) [33,34] calculation is based on the ratio of within-cluster distances to between-cluster distances.

$$DBI = \frac{1}{K} \sum_{i=1}^{i=K} R_i, \text{ where } R_{ij} = \frac{\bar{S}_i + \bar{S}_j}{d_{ij}}, \text{ and } S_i = \frac{1}{|C_i|} \sum_{h=1}^{h=|C_i|} d(x_{ih}, c_i) \quad (21)$$

$R_i = \max_{i \neq j} \{R_{ij}\}$ —maximum of  $R$  between the cluster  $C_i$  and each other cluster  $C_j$ ,

$d_{ij} = d(c_i, c_j)$ —distance between the centroid  $c_i$  of cluster  $C_i$  and centroid  $c_j$  of the cluster  $C_j$ ,

$\bar{S}_i$ —the average distance between every pixel (within the cluster  $C_i$ ) and its centroid  $c_i$ ,

$\bar{S}_j$ —the average distance between every pixel (within the cluster  $C_j$ ) and its centroid  $c_j$ .

Here, cohesion is defined in the form of sum  $S_i + S_j$ , while  $d_{ij}$  defines separation. Lower  $DBI$  values indicate a better clustering result. If clusters are close to each other (small  $d_{ij}$ ) and are dispersed (big  $\bar{S}_i + \bar{S}_j$ ), then  $DBI$  value will be high, indicating less optimal clustering.

### 5.3. IQA Metrics

IQA metrics can be divided into three groups, depending on the need of the reference information: (1) Full-Reference (FR)—metrics, that require reference/ground-truth image (quality of the segmented image is measured in comparison to the ground-truth image); (2) No Reference (NR)—metrics, that do not require reference/ground truth image for measuring quality; (3) reduced reference (RR) metrics measure quality by comparing distorted/segmented image with a reference/ground truth image, composed of specific extracted features (such as edge information). The reference image is named ground truth from the perspective of image segmentation. In this experiment, we concentrated on commonly used FR metrics that are listed in Table 6.

**Table 6.** Full-reference image quality assessment (FR-IQA) image quality metrics and their main features.

Notation	Ref <sup>1</sup>	Name	Metric Features <sup>2</sup>
<i>PSNR</i>	-	Peak Signal to Noise Ratio	pixel difference-based, inversely proportional to the <i>MSE</i>
<i>SR-SIM</i>	[35]	Spectral Residual based similarity	visual saliency map, gradient modulus
<i>UQI</i>	[36]	Universal quality index	structural distortion, luminance distortion, loss of contrast
<i>MS-SSIM</i>	[37]	Multi-scale Structural Similarity index	structural distortion
<i>IW-SSIM</i>	[38]	Information Content Weighted <i>SSIM</i>	NSS
<i>NQM</i>	[39]	Noise Quality Measure	CSF
<i>FSIMc</i>	[40]	Feature Similarity Index	structural distortion
<i>SSIM</i>	[41]	Structural Similarity	luminance, contrast, and structural distortions
<i>MAD</i>	[42]	Most Apparent Distortion	local luminance and contrast masking, spatial-frequency components changes, CSF
<i>GSM</i>	[43]	Gradient Similarity	luminance, structural and contrast changes
<i>WSNR</i>	[44]	Weighted signal to noise ratio	CSF
<i>SFF</i>	[45]	Sparse Feature Fidelity	based on sparse features similarity (structure differences) and luminance correlation (brightness distortions)
<i>VIF<sub>p</sub></i>	[46]	pixel-based Visual Information Fidelity	NSS
<i>VSNR</i>	[47]	Visual Signal-to-Noise Ratio	visual masking, perceived contrast, global precedence
<i>PSNR-HVS-M</i>	[48]	Extension of <i>PSNR</i>	incorporates CSF and between-coefficient contrast masking of DCT basis functions
<i>VSI</i>	[49]	Visual Saliency-induced index	visual saliency map, gradient modulus, colour distortions
<i>SNR</i>	-	Signal to Noise Ratio	Pixel difference-based

<sup>1</sup> All references contain research papers provided by the original authors of the specific quality method.

<sup>2</sup> NSS—Natural Scene Statistic models. CSF—Contrast Sensitivity Function. DCT—Discrete Cosine Transform. *MSE*—Mean Squared Error.

IQA metrics can be applied to the image segmentation evaluation [10]. As input, the segmented images have different characteristics compared to the loosely compressed natural images and can be treated more like a synthetic (i.e., artificially generated) image. Segmented images have crisp contours,

uniform regions, and are generally less complex. Majority of the proposed IQA metrics are designed for correlation with HVS, which is very sensitive to the contrast [50] or structural information changes. To achieve correlation with human perceived quality, most IQA metrics employ multiple strategies. However, some of the features (Table 6), like contrast changes, contrast masking or luminance masking, may be less or not important in evaluating the satellite image segmentation result. Various research studies emphasize the importance of precise contours for improved perceived segmentation quality, after combining *JI* and *BF* quality metrics [8] in the novel FR-IQA quality index, intended for colour images [5].

All the selected FR-IQA measures were calculated using MATLAB R2019b. *VSNR*, *VIF<sub>p</sub>*, *UQI*, *NQM*, and *WSNR* calculated from MeTriX MuX Visual Quality Assessment Package (v. 1.1) [51] using default settings.

## 6. Subjective Quality Assessment of Segmented Satellite Images

Subjective evaluation by a human is the most reliable method for determining image quality in various applications (such as image editing, image retargeting [52], and others) as well as in the image segmentation [7]. However, segmentation quality requirements are also application-dependent [8,9].

Subjective quality evaluation tests require careful preparation, human, and time resources. In contrast to the long-established subjective evaluation of the distorted image and video quality, image segmentation lacks dedicated quality evaluation methodologies. For this reason, due to similarities with IQA, many strategies for subjective quality assessment for segmentation are adopted from the existing standards [7].

For the assessment of image quality, there are many test methods and rating scales provided by International Telecommunication Union (ITU) standards, describing acceptable modifications and recommendations. The method describes how a stimulus (in this case, a sequence of segmented images) is presented, and the rating scale describes the way that subjects evaluate their opinion of the stimulus. For this subjective quality assessment test, the simultaneous double stimulus for continuous evaluation (SDSCE) method described in ITU-R BT.500-14 [53] was combined with an absolute category rating (ACR) scale described in T-REC-P.913 [54]. ACR consists of a five-level rating scale: (5—Excellent, 4—Good, 3—Fair, 2—Poor and 1—Bad). T-REC-P.913 also does not recommend increasing the number of levels, since the accuracy of the results does not improve [55], and the evaluation process becomes more complicated for a human.

Before performing the *MOS* experiment, it is recommended to ensure that there are enough segmentation results of bad, average, and good quality between all segmented images. Otherwise, the data points will be concentrated in a single corner of the scatter plot and will not fully cover the ACR scale.

For subjects to fully understand their task, the first part of the test is the training phase and includes examples covering the full range of segmentation quality results combined with verbal instructions given by the administrator explaining voting procedure by following practices described in Reference (Section 11.5 in [54]).

During the second phase, subjects were presented with the electronic form depicted in Figure 4 and were requested to evaluate the differences between the GT segmentation and segmentation result obtained by *k-means++*. Subjects were aware of which image is the GT and which image is the segmentation result obtained by the selected algorithm. There was no time limit for evaluating a single pair of images. However, each experiment session lasted no more than 20 min.

Similar to the approach in Reference [9], original satellite images were also placed on the left, but only for context purposes and should not impact the judgement of segmentation quality. The segmented images have not been scaled to avoid introducing possible distortions.

In order to avoid fatigue and/or boredom, the test was divided into two parts with each consisting of 45 segmentations (90 segmentations in total). The first part was evaluated by 95 and the second part was evaluated by 92 subjects, which is more than enough for stable average ratings.

**Figure 4.** Example of the electronic form for segmentation quality assessment using the absolute category rating (ACR) scale. Different image clusters are marked with selected colours in “Ground truth” and “Segmentation result” images.

After collecting experimental results, ratings were converted to numerical values (1, 2, 3, 4, and 5) and the total *MOS* for the single-segmented image was calculated as the arithmetic mean of the individual assessments that the subjects assigned to the segmentation result [53].

$$MOS = \frac{1}{N} \cdot \sum_{n=1}^N o_n \quad (22)$$

Here,  $o_n$  is the observed rating for subject  $n$ , and  $N$  is the total number of subjects (participants) in the experiment. We observed that most segmented images received *MOS* scores from 2.5 to 3.5. This distribution is likely due to the human tendency of trying to avoid giving extreme scores when evaluating images [56].

## 7. Results

The workflow presented in Figure 5 was performed to collect all necessary data required for calculating the correlation between the subjective and objective scores.

As seen in the general framework, the whole workflow can be divided into three main branches. On the left section are presented steps that deal with constructing our dataset using the original dataset of satellite images and their GT. In the middle section are the described steps related to the selection of the segmentation method and performing segmentation of the satellite images based on the colour feature. Presented steps that are used for obtaining subjective scores are on the right section. The goal is to obtain the objective and subjective scores for the calculation of the Pearson Linear Correlation Coefficient (*PLCC*) and Spearman Rank Order Correlation Coefficient (*SROCC*), which is the final step and is further described in this section. The calculation of the objective scores is presented before the final step.

Correlation between the subjective (*sub*) and objective (*obj*) scores was determined by *PLCC* (23) and *SROCC* (24) [57]. Here,  $n$  is the total number of images in the dataset.

$$PLCC = \frac{n(\sum sub \cdot obj) - (\sum sub)(\sum obj)}{\sqrt{[n \sum sub^2 - (\sum sub)^2] \cdot [n \sum obj^2 - (\sum obj)^2]}} \quad (23)$$

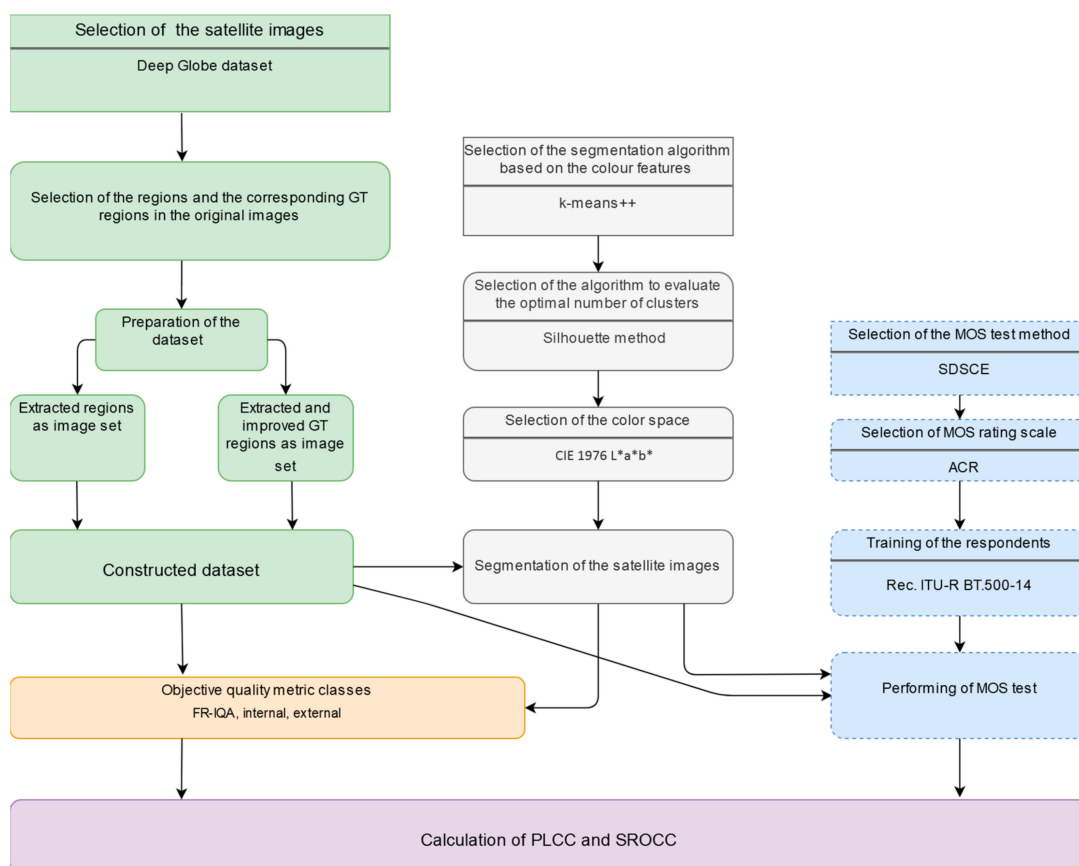
*SROCC* is equal to *PLCC* applied on the ranks, in this case, of subjective and objective scores. For the image  $i$ ,  $d_i$  denotes the difference between the ranks. In order to compute *SROCC*, data have to be converted into ranks. Assuming no tied ranks exist, a simplified *SROCC* formula can be applied.

$$SROCC = 1 - \frac{6 \cdot \sum d_i^2}{n(n^2 - 1)}. \quad (24)$$

*PLCC* is the measurement of the strength of the linear relationship, while *SROCC* also measures the strength of the monotonic relationship. For example, for images distorted by various compression artifacts, most IQA metrics display nonlinear relationships with HVS [58]. Both correlation coefficients have a  $[-1; 1]$  range. Negative values indicate a negative correlation, while positive values indicate a positive correlation. If the values of both correlation coefficients are approaching zero, this suggests that relationship most likely does not exist.

The results of the experiment are presented in the two groups of Tables: Tables 7–9 for the overall correlation between subjective *MOS* scores and objective metric scores and Tables 10–12 for the correlation between different quality groups of *MOS* scores and objective metric scores.

Tables 7–9 show the overall correlation using all images from the dataset. In Tables 8 and 9, metrics are sorted from highest to lowest, according to their *PLCC* and *SROCC* scores. Table 7 shows that the *SH* has the strongest positive correlation, while the *DBI* has the strongest negative correlation for both *PLCC* and *SROCC*. Table 8 presents that the *Jl* has the highest *PLCC*, while *ACC* has the highest *SROCC* closely followed by *Jl*. Table 9 shows that the *PSNR* has the highest *PLCC*, while *UQI* has the highest *SROCC* closely followed by *PSNR*.



**Figure 5.** Framework illustrating the workflow for obtaining final results (different colours present different branches of the workflow, including the appropriate steps).

**Table 7.** The overall correlation between subjective MOS scores and internal validation scores.

Metric <sup>1</sup>	PLCC	SROCC	Remarks
Silhouette coefficient ( <i>SH</i> )	0.4684	0.4252	Higher is better
Calinski–Harabasz index ( <i>CHI</i> )	0.3902	0.3426	
<i>SSB</i>	0.1861	0.1636	
Hartigan index ( <i>HI</i> )	0.0443	0.0428	
Xu coefficient ( <i>Xu</i> )	0.0173	0.0153	Lower is better
<i>SSE</i>	0.1477	0.1698	
Davies–Bouldin index ( <i>DBI</i> )	−0.4562	−0.4355	

<sup>1</sup> Internal metrics were calculated using the optimal number of clusters selected by the Silhouette method.

**Table 8.** The overall correlation between subjective MOS scores and external validation scores.

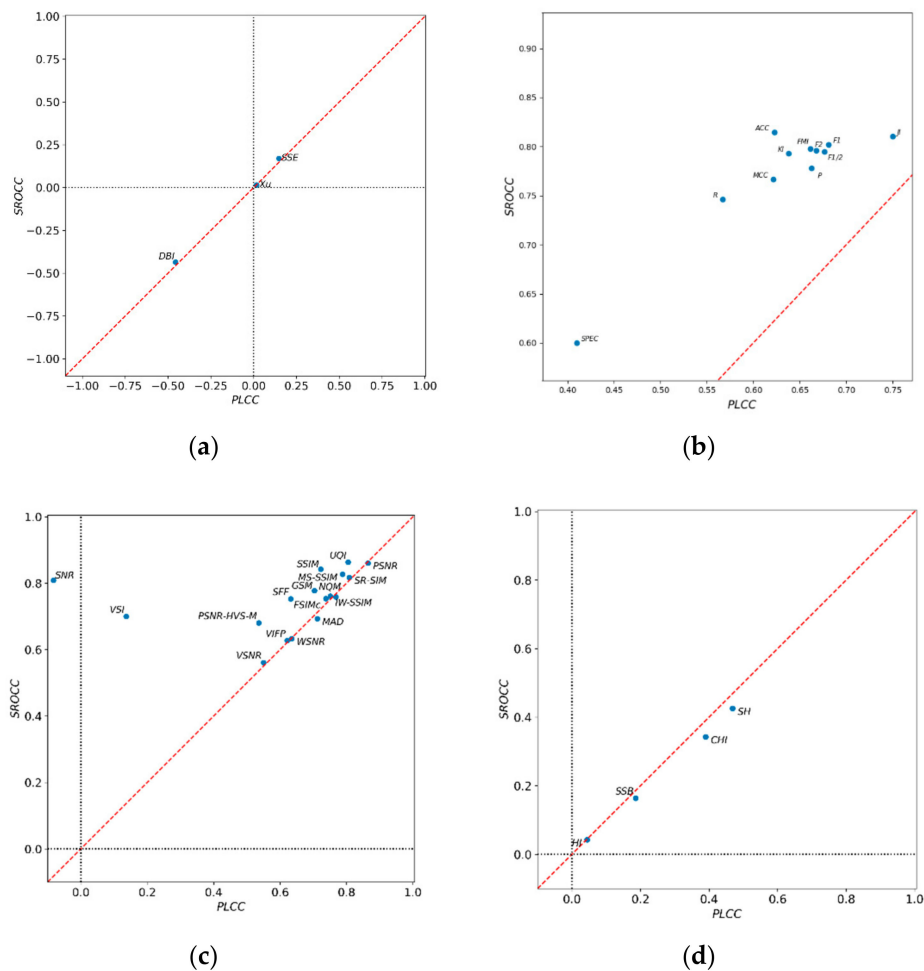
Metric	PLCC	Metric	SROCC
Jaccard index ( <i>JI</i> )	0.7497	Accuracy ( <i>ACC</i> )	0.8147
Fscore ( <i>F<sub>1</sub></i> )	0.681	Jaccard index ( <i>JI</i> )	0.8105
<i>F<sub>1/2</sub></i>	0.6764	Fscore ( <i>F<sub>1</sub></i> )	0.802
Fowlkes–Mallows index ( <i>FMI</i> )	0.6673	<i>F<sub>2</sub></i>	0.7981
Precision ( <i>P</i> )	0.6623	Fowlkes–Mallows index ( <i>FMI</i> )	0.7961
<i>F<sub>2</sub></i>	0.6613	<i>F<sub>1/2</sub></i>	0.7947
Kulczynski index ( <i>KI</i> )	0.6379	Kulczynski index ( <i>KI</i> )	0.7928
Accuracy ( <i>ACC</i> )	0.6224	Precision ( <i>P</i> )	0.7782
<i>MCC</i>	0.6214	<i>MCC</i>	0.7665
Recall/Sensitivity ( <i>R</i> )	0.5668	Recall/Sensitivity ( <i>R</i> )	0.7464
Specificity ( <i>SPEC</i> )	0.4096	Specificity ( <i>SPEC</i> )	0.5998

**Table 9.** The overall correlation between subjective MOS scores and FR-IQA measures.

Metric	PLCC	Metric	SROCC
<i>PSNR</i>	0.8647	<i>UQI</i>	0.8632
<i>SR-SIM</i>	0.8083	<i>PSNR</i>	0.8608
<i>UQI</i>	0.805	<i>SSIM</i>	0.8423
<i>MS-SSIM</i>	0.7881	<i>MS-SSIM</i>	0.8264
<i>NQM</i>	0.7677	<i>SR-SIM</i>	0.8174
<i>IW-SSIM</i>	0.7511	<i>SNR</i>	0.8087
<i>FSIMc</i>	0.7381	<i>GSM</i>	0.7774
<i>SSIM</i>	0.7226	<i>IW-SSIM</i>	0.761
<i>MAD</i>	0.7118	<i>NQM</i>	0.7586
<i>GSM</i>	0.7036	<i>FSIMc</i>	0.7534
<i>WSNR</i>	0.6351	<i>SFF</i>	0.753
<i>SFF</i>	0.632	<i>VSI</i>	0.7001
<i>VIF<sub>p</sub></i>	0.6214	<i>MAD</i>	0.6932
<i>VSNR</i>	0.5501	<i>PSNR-HVS-M</i>	0.6802
<i>PSNR-HVS-M</i>	0.5357	<i>WSNR</i>	0.6333
<i>VSI</i>	0.1365	<i>VIFP</i>	0.6276
<i>SNR</i>	−0.082	<i>VSNR</i>	0.5609

Information from Tables 7–9 is also represented as the scatterplots in Figure 6a–d, respectively, showing a comparison between *PLCC* and *SROCC* values. Vertical and horizontal axes correspond to *SROCC* and *PLCC*, respectively. Here, metrics close to the dotted red line ( $y = x$ ) have similar *PLCC* and *SROCC* values. Due to very similar values for external metrics, the scatterplot scale and position in Figure 6b were adjusted. The closer metric is to the (0, 0) point, the weaker correlation with *MOS* for that metric is. Best results are for metrics, which are closer to the upper right corner (positive correlation with *MOS*) or lower-left corner (negative correlation with *MOS*). From the visual inspection, it can also be observed that *SROCC* and *PLCC* values themselves display strong positive linear correlation, meaning both correlation coefficients are equally important.

Tables 10–12 show the correlation between each metric and different quality groups based on MOS: low-quality (1.0–2.5), middle-quality (2.5–3.5), and high-quality (3.5–5]. For each quality group, the top three best results are highlighted except the results of the correlation in Table 10. Additionally, the information in Tables 10–12 are also presented as the bar charts in Figures 7–9. For each quality group in Tables 11 and 12, the top three best results are highlighted in light green.



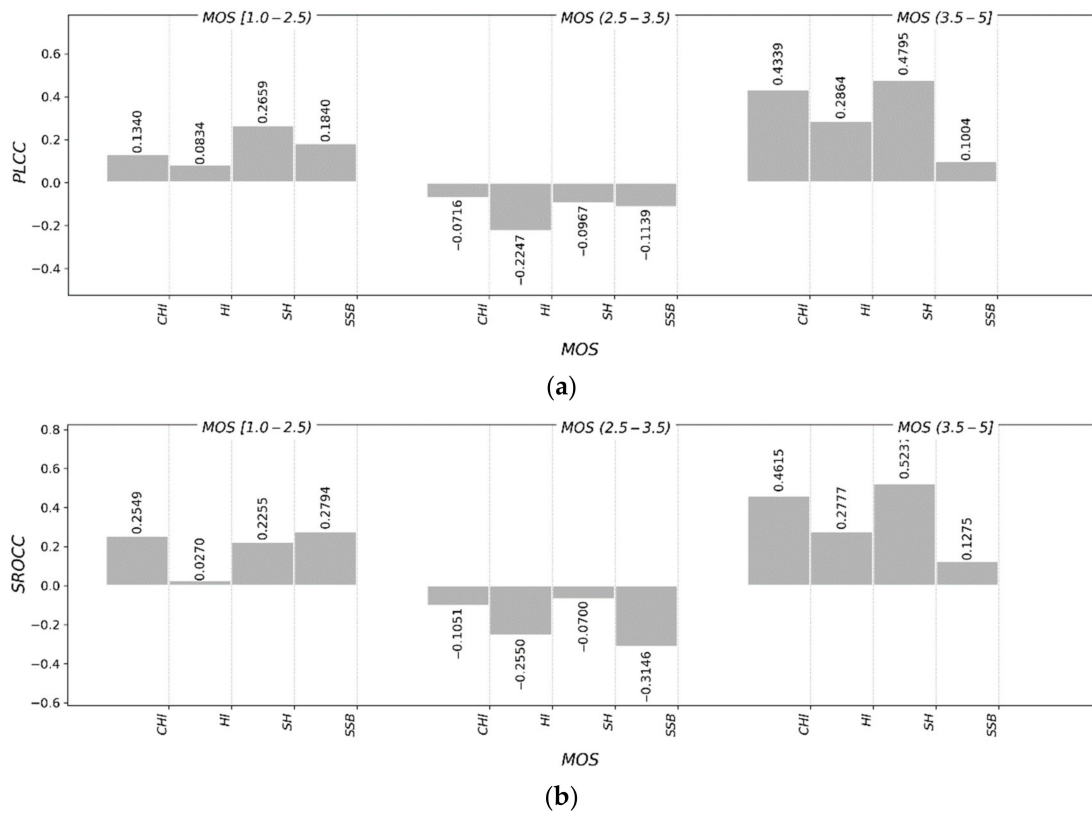
**Figure 6.** Comparison of SROCC and PLCC values (obtained from Tables 7–9) for: (a) internal validation scores (DBI, Xu, SSE), (b) external validation scores, (c) FR-IQA measures, and (d) internal validation scores (HI, SSB, CHI, SH). Red  $y=x$  line denotes where SROCC and PLCC values are equal.

**Table 10.** Correlation between different quality groups of MOS scores and internal validation scores.

Metric <sup>1</sup>	PLCC			SROCC			Remarks	
	MOS	[1.0–2.5)	(2.5–3.5)	(3.5–5]	[1.0–2.5)	(2.5–3.5)		(3.5–5]
Calinski–Harabasz index (CHI)		0.1340	−0.0716	0.4339	0.2549	−0.1051	0.4615	Higher is better
SSB		0.1840	−0.1139	0.1004	0.2794	−0.3146	0.1275	
Hartigan index (HI)		0.0834	−0.2247	0.2864	0.0270	−0.2550	0.2777	
Silhouette coefficient (SH)		0.2659	−0.0967	0.4795	0.2255	−0.0700	0.5237	Lower is better
Davies–Bouldin index (DBI)		−0.1798	0.0126	−0.5035	−0.1104	−0.0057	−0.5198	
SSE		0.2734	−0.0988	−0.3966	0.5319	−0.0757	−0.3706	
Xu coefficient (Xu)		0.2881	−0.1576	−0.4714	0.4583	−0.1436	−0.4684	

<sup>1</sup> Internal metrics were calculated using the optimal number of clusters selected by the Silhouette method.

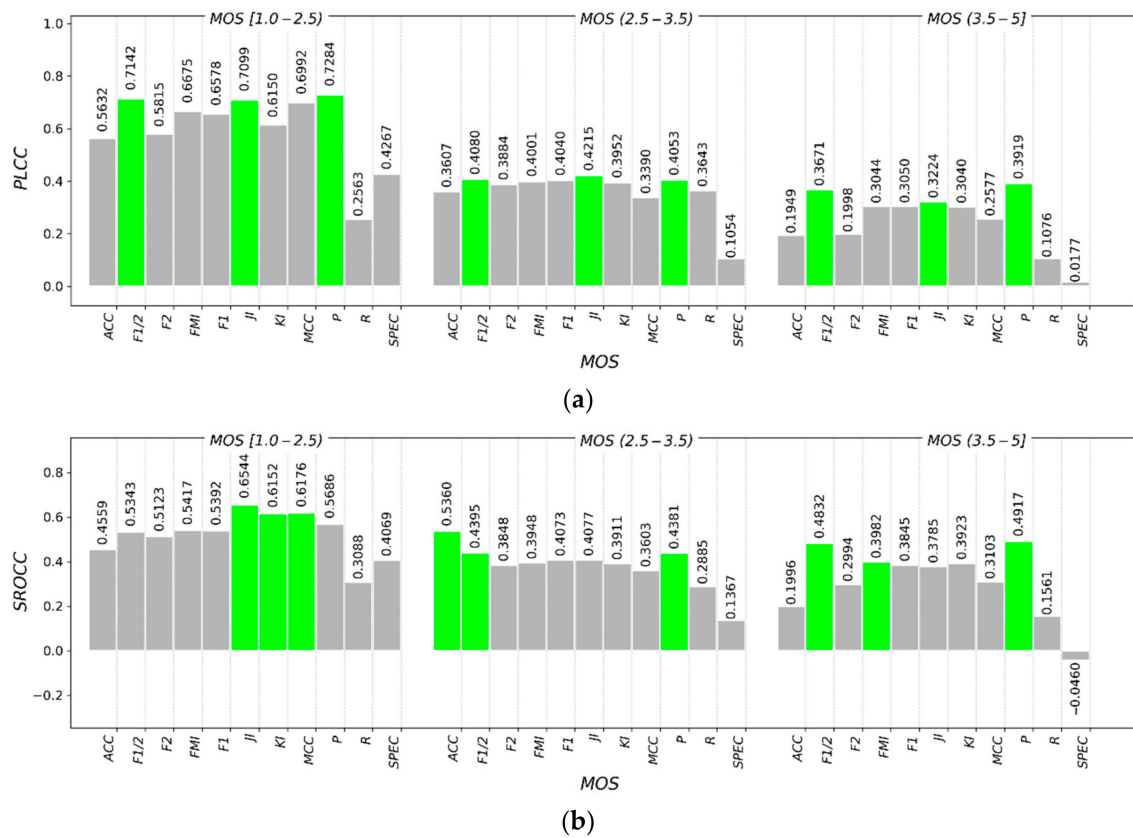




**Figure 7.** Comparison of the correlation values between different quality groups of MOS scores and internal validation scores (CHI, HI, SH, SSB): (a) PLCC, and (b) SROCC.

**Table 11.** Correlation between different quality groups of MOS scores and external validation scores (the top three best results are highlighted for each quality group).

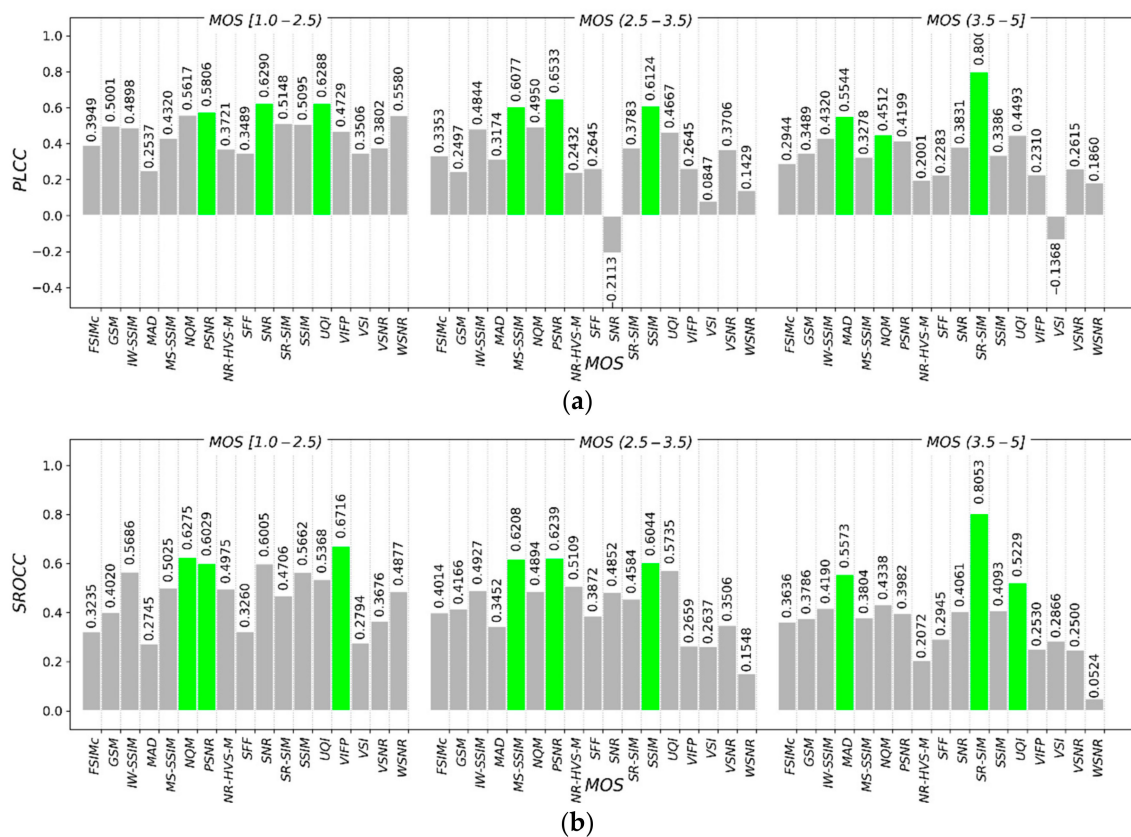
Metric	PLCC			SROCC			
	MOS	[1.0–2.5]	(2.5–3.5)	(3.5–5]	[1.0–2.5]	(2.5–3.5)	(3.5–5]
Accuracy (ACC)		0.5632	0.3607	0.1949	0.4559	<b>0.5360</b>	0.1996
$F_{1/2}$		<b>0.7142</b>	<b>0.4080</b>	<b>0.3671</b>	0.5343	<b>0.4395</b>	<b>0.4832</b>
$F_2$		0.5815	0.3884	0.1998	0.5123	0.3848	0.2994
Fowlkes–Mallows index (FMI)		0.6675	0.4001	0.3044	0.5417	0.3948	<b>0.3982</b>
Fscore ( $F_1$ )		0.6578	0.4040	0.3050	0.5392	0.4073	0.3845
Jaccard index (JI)		<b>0.7099</b>	<b>0.4215</b>	<b>0.3224</b>	<b>0.6544</b>	0.4077	0.3785
Kulczynski index (KI)		0.6150	0.3952	0.3040	<b>0.6152</b>	0.3911	0.3923
MCC		0.6992	0.3390	0.2577	<b>0.6176</b>	0.3603	0.3103
Precision (P)		<b>0.7284</b>	<b>0.4053</b>	<b>0.3919</b>	0.5686	<b>0.4381</b>	<b>0.4917</b>
Recall/Sensitivity (R)		0.2563	0.3643	0.1076	0.3088	0.2885	0.1561
Specificity (SPEC)		0.4267	0.1054	0.0177	0.4069	0.1367	−0.0460



**Figure 8.** Comparison of the correlation values between different quality groups of MOS scores and external validation scores (the top three best results are highlighted in light green for each quality group): (a) PLCC, and (b) SROCC.

**Table 12.** Correlation between different quality groups of MOS scores and FR-IQA measures (the top three best results are highlighted for each quality group).

Metric	PLCC			SROCC		
	MOS [1.0-2.5]	MOS [2.5-3.5]	MOS [3.5-5]	MOS [1.0-2.5]	MOS [2.5-3.5]	MOS [3.5-5]
FSIM <sub>c</sub>	0.3949	0.3353	0.2944	0.3235	0.4014	0.3636
GSM	0.5001	0.2497	0.3489	0.4020	0.4166	0.3786
IW-SSIM	0.4898	0.4844	0.4320	0.5686	0.4927	0.4190
MAD	0.2537	0.3174	<b>0.5544</b>	0.2745	0.3452	<b>0.5573</b>
MS-SSIM	0.4320	<b>0.6077</b>	0.3278	0.5025	<b>0.6208</b>	0.3804
NQM	0.5617	0.4950	<b>0.4512</b>	<b>0.6275</b>	0.4894	0.4338
PSNR	<b>0.5806</b>	<b>0.6533</b>	0.4199	<b>0.6029</b>	<b>0.6239</b>	0.3982
PSNR-HVS-M	0.3721	0.2432	0.2001	0.4975	0.5109	0.2072
SFF	0.3489	0.2645	0.2283	0.3260	0.3872	0.2945
SNR	<b>0.6290</b>	-0.2113	0.3831	0.6005	0.4852	0.4061
SR-SIM	0.5148	0.3783	<b>0.8001</b>	0.4706	0.4584	<b>0.8053</b>
SSIM	0.5095	<b>0.6124</b>	0.3386	0.5662	<b>0.6044</b>	0.4093
UQI	<b>0.6288</b>	0.4667	0.4493	0.5368	0.5735	<b>0.5229</b>
VIF <sub>P</sub>	0.4729	0.2645	0.2310	<b>0.6716</b>	0.2659	0.2530
VSI	0.3506	0.0847	-0.1368	0.2794	0.2637	0.2866
VSNR	0.3802	0.3706	0.2615	0.3676	0.3506	0.2500
WSNR	0.5580	0.1429	0.1860	0.4877	0.1548	0.0524



**Figure 9.** Comparison of the correlation values between different quality groups of MOS scores and FR-IQA measures (the top three best results are highlighted in light green for each quality group): (a) PLCC, and (b) SROCC.

Dividing results into three different quality groups allows for the more specialized comparison. The global correlation scores computed for the full dataset (Tables 7–9) do not allow identifying metrics that have a moderate correlation for all images (regardless of segmentation quality) from metrics that have a high correlation for images with above-average segmentation quality and low correlation for images with bad segmentation quality.

### 8. Discussion

The internal measures do not correlate well with perceived segmentation quality (the overall results in Table 7 and Figure 6a,d comparing to the external Table 8 and Figure 6b and FR-IQA measures (Table 9 and Figure 6c). Although the internal measures are used to evaluate results of clustering algorithms and help to select an optimal number of clusters, they were never intended to correlate with perceived image segmentation quality and relates more to how k-means is optimized. Thus, poor correlation is to be expected. In contrast, internal metrics combining cohesion and separation (such as *DBI*, *CHI*, and *SH*) seems to achieve slightly higher correlations, but as in Table 10 and Figure 7, only for segmentation results with high MOS scores. Note *DBI* shows a negative correlation as this measure returns lower scores for better clustering solutions. Overall, *SH* achieved moderate correlation for images in the MOS range (3.5–5).

The overall correlation scores for the external validation metrics are in a tight cluster (Table 8 and Figure 6b). As presented in Table 11 and Figure 8, lower MOS values correspond to the stronger correlations. This could be explained as the respondents agreeing more when determining the lower overall quality of the segmented image [7], while the opinions for better segmentation quality differs more. For the images in the low-quality group, *P*,  $F_{1/2}$ , and *JI* show a strong positive linear correlation,

$MOS > 0.7$ . The  $F_{1/2}$  and  $P$  metrics share a very similar correlation, as  $F_{1/2}$  has a higher weight for  $P$ . Since there were no extended studies in this area, results in their entirety cannot be compared to previous studies. However, we observed  $SROCC$  values obtained by authors [7] for  $JI$  and  $F_1$  metrics (0.848 and 0.848) used for single object segmentation are close to our overall  $SROCC$  values (0.811 and 0.802).

Best overall correlation for the IQA metric group was achieved by  $PSNR$ ,  $UQI$ ,  $SSIM$ , and  $SR-SIM$  ( $MOS > 0.8$ ) (Table 9 and Figure 6c). All of them are also known to be very fast, according to the average calculation time for natural images from the TID2013 database [59]. These could be a reasonable choice for evaluating segmentation quality in terms of calculation speed and correlation with human perception.

It is widely accepted that the  $PSNR$  does not always agree with HVS assessing the quality of natural images distorted by various compression methods. From the overall results in Table 9 and Figure 6c,  $PSNR$  does not hold a significant advantage in the  $MOS$  score over most of other IQA metrics. Compared in the quality groups (Table 12 and Figure 9), we can state the similarity of  $IW-SSIM$  to  $PSNR$ .  $IW-SSIM$  is a relatively stable metric according to  $PLCC$ , and  $PSNR$  is stable only in the low-middle  $MOS$  (1.0–3.5) segmentation quality. The  $UQI$  metric is the most stable according to  $SROCC$ , while  $PSNR$  is only performing better for the low-middle segmentation quality and may be not an optimal choice depending on the segmentation situation.  $SR-SIM$  is a better choice for high-quality segmentations,  $MOS$  (3.5–5].

The reason for  $PSNR$  being able to compete with the advanced HVS metrics could be the nature of the segmented images, differing from the natural ones. The most distinct features of segmented images are clear contours and object shapes, uniform regions and colour, and absence of noise. HVS is sensitive to the structural information changes in the images. Therefore, metrics like  $UQI$  and  $SSIM$  are effective. HVS also largely relies on edge information for image interpretation [60]. In Reference [10], authors state that  $PSNR$  can be a good method to evaluate edge detection algorithms (for example, in BSR300 database).

The IQA metrics, in general, have a varying correlation to subjective assessment depending on the database, distortion type, image content, and image segmentation quality categories.

Authors of Reference [60] state the  $PSNR$  can sometimes have a strong correlation: 0.8756 ( $SROCC$ ) and 0.8723 ( $PLCC$ ) using natural images from the LIVE database. These results are very similar to our overall results: 0.8647 ( $PLCC$ ) and 0.8608 ( $SROCC$ ).  $PSNR$  does not outperform HVS-based metrics, which tend to be more stable across different databases of natural images and have even higher performance scores ( $MOS$  score).

Possible future directions for this research can include additional testing for RR, NR-IQA relations to subjective evaluation using larger-scale segmentation dataset(s) including multispectral satellite images. We suppose it is possible to select the prominent metrics for the combined measure, designed for improved correlation with subjective quality scores of the segmented images. To expand further, it can be interesting to evaluate the influence of algorithm selection, like using DBSCAN or another algorithm instead of k-means++, to the impact of correlation scores.

Depending on the goal or method, obtained segmentation results may have clusters assigned to a different colour value. For example, a segmented image consists of shades of a single colour versus very contrasting (or opposite) colour segments. Higher colour differences may impact some FR-IQA metric results and/or subjective evaluation itself. However, humans are able to distinguish more different levels of colour shades compared to gray shades [5]. Finally, determining a correlation among different groups of metrics might provide additional insight and a more diverse comparison.

## 9. Conclusions

This research aims to evaluate the correlation between the subjective and objective image quality metrics from the perspective of satellite image segmentation. Three broad classes of quality metrics were considered: internal and external cluster validation indices as well as FR-IQA measures.

From state-of-the-art technologies, we can conclude that there is no extensive research related to the assessment of the effectiveness of satellite image segmentation. For the segmentation quality test, we constructed our dataset of the satellite images with GT, based on “DeepGlobe Land Cover Classification Challenge” dataset.

From the experimental studies, several essential observations related to the assessment of the effectiveness of satellite image segmentation are made.

- When the segmentation results are diverse in perceived quality, then most external measures and FR-IQA metrics display very similar correlation with MOS.
- As perceived segmentation quality decreases, the correlation with MOS increases for the external quality measures.
- The PSNR metric achieved consistent results for low-middle quality segmentation, MOS range [1.0–3.5).
- The best metric for evaluating high-quality segmentation (MOS range (3.5–5)) was SR-SIM, achieving SROCC, and PLCC scores above 0.8, which also have low computational complexity.
- Since PSNR and SR-SIM complement each other covering full MOS range, they could be combined into a single measure.

The experimental studies show that dividing segmentation results into three different quality groups based on MOS allows the more specialized comparison of the objective quality metrics, according to perceived image quality.

Our study might provide insights to other research, where selecting the most suitable subjective metric from the HVS perspective is crucial. Herewith, our original results and obtained observations can be applied for improving current state-of-the-art segmentation methods.

**Author Contributions:** Conceptualization, G.K.-J. and E.J. Methodology, G.K.-J., E.J., R.B., T.L., and M.K. Software, E.J. Validation, G.K.-J., E.J., R.B., T.L., and M.K. Formal analysis, G.K.-J., E.J., R.B., T.L., and M.K. Investigation, G.K.-J. and E.J. Resources, E.J. Data curation, G.K.-J. and E.J. Writing—original draft preparation, E.J. Writing—review and editing, G.K.-J., E.J., R.B., T.L., and M.K. Supervision, G.K.-J., R.B., T.L., and M.K. Project administration, G.K.-J., R.B., T.L., and M.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-MIP-19-27.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deep Globe 2018: A Challenge to Parse the Earth through Satellite Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–17209.
2. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2217–2226. [[CrossRef](#)]
3. Chitade, A.; Katiyar, S. Color Based Image Segmentation Using K-Means Clustering. *Int. J. Eng. Sci. Technol.* **2010**, *2*, 5319–5325.
4. Chen, H.; Wang, S. Visible Color Difference-Based Quantitative Evaluation of Color Segmentation. *Vis. Image Signal Process. IEE Proc.* **2006**, *153*, 598–609. [[CrossRef](#)]
5. Gupta, P.; Srivastava, P.; Bhardwaj, S.; Bhateja, V.A. *Novel Full-Reference Image Quality Index for Color Images*; InConINDIA 2012, AISC 132; Satapathy, S.C., Avadhani, P.S., Abraham, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 245–253.
6. Egiazarian, K.; Astola, J.; Lukin, B.; Battisti, F.; Carli, M. *A New Full-Reference Quality Metrics Based on HVS*; Semantic Scholar: Seattle, WA, USA, 2006.
7. Shi, R.; Ngan, K.; Li, S.; Paramesran, R.; Li, H. Visual Quality Evaluation of Image Object Segmentation: Subjective Assessment and Objective Measure. *IEEE Trans. Image Process.* **2015**, *24*, 5033–5045. [[CrossRef](#)]

8. Csurka, G.; Larlus, D.; Perronnin, F. What is a Good Evaluation Measure for Semantic Segmentation? In Proceedings of the 24th British Machine Vision Conference, Bristol, UK, 9–13 September 2013.
9. Shi, R.; Ngan, K.N.; Li, S. Jaccard Index Compensation for Object Segmentation Evaluation. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 4457–4461.
10. Fardo, F.A.; Conforto, V.H.; Oliveira, F.C.; Rodrigues, P.S. A Formal Evaluation of PSNR as Quality Measurement Parameter for Image Segmentation Algorithms. *Comput. Vis. Pattern Recognit.* **2016**. ArXiv:1605.07116.
11. Marçal, A.R.; Rodrigues, A.; Cunha, M. Evaluation of Satellite Image Segmentation Using Synthetic Images. In Proceedings of the 2010 IEEE International Geoscience and Remote Sensing Symposium, Honolulu, HI, USA, 25–30 July 2010; pp. 2210–2213.
12. Wu, M.; Zhang, C.; Liu, J.; Zhou, L.; Li, X. Towards Accurate High Resolution Satellite Image Semantic Segmentation. *IEEE Access.* **2019**, *7*, 55609–55619. [[CrossRef](#)]
13. Singha, M.; Hemachandran, K. Color Image Segmentation for Satellite Images. *Int. J. Comput. Sci. Eng.* **2011**, *3*, 3756–3762.
14. Sirmaçek, B.; Unsalan, C. Road Detection from Aerial Images Using Color Features. In Proceedings of the 5th International Conference on Recent Advances in Space Technologies—RAST2011, Istanbul, Turkey, 9–11 June 2011.
15. Al-Ghraiiri, A.; Abed, Z.H.; Fadhil, F.; Naser, F.K. Classification of Satellite Images Based on Color Features Using Remote Sensing. *Int. J. Comput. IJC* **2018**, *31*, 42–52.
16. Silva, R.D.; Minetto, R.; Schwartz, W.; Pedrini, H. Satellite Image Segmentation Using Wavelet Transforms Based on Color and Texture Features. *ISVC 2008, Part II*, 113–122.
17. MacQueen, J.B. 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: Berkeley, CA, USA, 1967; pp. 281–297.
18. Fränti, P.; Sieranoja, S. How much can k-means be improved by using better initialization and repeats? *Pattern Recognit.* **2019**, *93*, 95–112. [[CrossRef](#)]
19. Rousseeuw, P. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [[CrossRef](#)]
20. Arthur, D.; Vassilvitskii, S. K-Means++: The Advantages of Careful Seeding. In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA) 2007, New Orleans, LA, USA, 7–9 January 2007.
21. Moghaddam, S.Z.; Monadjemi, A.; Nematbakhsh, N. Color Image Segmentation using Multi-thresholding Histogram and Morphology. *Int. J. Res. Rev. Comput. Sci.* **2012**, *3*, 1576–1579.
22. Martin, D.; Fowlkes, C.C.; Tal, D.; Malik, J. A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In Proceedings of the Eighth IEEE International Conference on Computer Vision, ICCV, Vancouver, BC, Canada, 7–14 July 2001; Volume 2, pp. 416–423.
23. The Assessment of the Segmentation Effectiveness of the Satellite Images. Available online: [https://drive.google.com/drive/folders/10SqFSSCiUOA2gbJ\\_Y2l5gGhUcEoEJeej?usp=sharing](https://drive.google.com/drive/folders/10SqFSSCiUOA2gbJ_Y2l5gGhUcEoEJeej?usp=sharing) (accessed on 29 October 2020).
24. Palacio-Niño, J.; Galiano, F. Evaluation Metrics for Unsupervised Learning Algorithms. *arXiv* **2019**. arxiv:905.05667.
25. Wang, Z.; Wang, E.; Zhu, Y. Image Segmentation Evaluation: A survey of Methods. *Artif. Intell. Rev.* **2020**, *53*, 5637–5674. [[CrossRef](#)]
26. Zhang, H.; Fritts, J.; Goldman, S.A. Image Segmentation Evaluation: A Survey of Unsupervised Methods. *Comput. Vis. Image Underst.* **2008**, *110*, 260–280. [[CrossRef](#)]
27. Desgraupes, B. *Clustering Indices*; University Paris Ouest. Lab Modal’X: Nanterre, France, 2016.
28. MATLAB Central File Exchange. Available online: <https://www.mathworks.com/matlabcentral/fileexchange/46035-confusionmatstats-group-grouphat> (accessed on 29 October 2020).
29. Tan, P.N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*; AddisonWesley: Boston, MA, USA, 2005.
30. Calinski, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat. Theory Methods* **1974**, *3*, 1–27. [[CrossRef](#)]

31. Hartigan, J.A. *Clustering Algorithms. Probability and Mathematical Statistics*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1975.
32. Xu, L. Bayesian Ying-Yang Machine, Clustering and Number of Clusters. *Pattern Recognit. Lett.* **1997**, *18*, 1167–1178. [[CrossRef](#)]
33. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227. [[CrossRef](#)]
34. Gustriansyah, R.; Suhandi, N.; Antony, F. Clustering Optimization in RFM Analysis Based on K-Means. *Indones. J. Electr. Eng. Comput. Sci.* **2020**, *18*, 470–477. [[CrossRef](#)]
35. Zhang, L.; Li, H. SR-SIM: A Fast and High Performance IQA Index Based on Spectral Residual. In Proceedings of the 2012 19th IEEE International Conference on Image Processing, Lake Buena Vista, Orlando, FL, USA, 30 September–3 October 2012; pp. 1473–1476.
36. Wang, Z.; Bovik, A. A Universal Image Quality Index. *IEEE Signal Process. Lett.* **2002**, *9*, 81–84. [[CrossRef](#)]
37. Wang, Z.; Simoncelli, E.P.; Bovik, A. Multiscale Structural Similarity for Image Quality Assessment. In Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 9–12 November 2003; Volume 2, pp. 1398–1402.
38. Wang, Z.; Li, Q. Information Content Weighting for Perceptual Image Quality Assessment. *IEEE Trans. Image Process.* **2011**, *20*, 1185–1198. [[CrossRef](#)] [[PubMed](#)]
39. Damera-Venkata, N.; Kite, T.; Geisler, W.; Evans, B.; Bovik, A. Image Quality Assessment Based on a Degradation Model. A publication of the IEEE Signal Processing Society. *IEEE Trans. Image Process.* **2000**, *9*, 636–650. [[CrossRef](#)] [[PubMed](#)]
40. Zhang, L.; Zhang, L.; Mou, X.; Zhang, D. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Trans. Image Process.* **2011**, *20*, 2378–2386. [[CrossRef](#)] [[PubMed](#)]
41. Wang, Z.; Bovik, A.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
42. Larson, E.C.; Chandler, D. Most Apparent Distortion: Full-Reference Image Quality Assessment and the Role of Strategy. *J. Electron. Imaging* **2010**, *19*, 011006.
43. Liu, A.; Lin, W.; Narwaria, M. Image Quality Assessment Based on Gradient Similarity. *IEEE Trans. Image Process.* **2012**, *21*, 1500–1512.
44. Mitsa, T.; Varkur, K.L. Evaluation of Contrast Sensitivity Functions for the Formulation of Quality Measures Incorporated in Halftoning Algorithms. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, MN, USA, 27–30 April 1993; Volume 5, pp. 301–304.
45. Chang, H.; Yang, H.; Gan, Y.; Wang, M. Sparse Feature Fidelity for Perceptual Image Quality Assessment. *IEEE Trans. Image Process.* **2013**, *22*, 4007–4018. [[CrossRef](#)]
46. Sheikh, H.; Bovik, A. Image Information and Visual Quality. *IEEE Trans. Image Process.* **2006**, *15*, 430–444. [[CrossRef](#)]
47. Chandler, D.; Hemami, S. VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images. *IEEE Trans. Image Process.* **2007**, *16*, 2284–2298. [[CrossRef](#)]
48. Ponomarenko, N.; Silvestri, F.; Egiazarian, K.; Carli, M.; Astola, J.; Lukin, V. On Between-Coefficient Contrast Masking of DCT Basis Functions, CD-ROM. In Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-07, Scottsdale, AZ, USA, 25–26 January 2007; p. 4.
49. Zhang, L.; Shen, Y.; Li, H. VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment. *IEEE Trans. Image Process.* **2014**, *23*, 4270–4281.
50. Min, X.; Gu, K.; Zhai, G.; Hu, M.; Yang, X. Saliency-Induced Reduced-Reference Quality Index for Natural Scene and Screen Content Images. *Signal Process.* **2018**, *145*, 127–136. [[CrossRef](#)]
51. MeTriX MuX Visual Quality Assessment Package. Available online: [https://github.com/sattarab/image-quality-tools/tree/master/metrix\\_mux](https://github.com/sattarab/image-quality-tools/tree/master/metrix_mux) (accessed on 29 October 2020).
52. Ma, L.; Lin, W.; Deng, C.; Ngan, K.N. Image Retargeting Quality Assessment: A Study of Subjective Scores and Objective Metrics. *IEEE J. Sel. Top. Signal Process.* **2012**, *6*, 626–639. [[CrossRef](#)]
53. International Telecommunication Union (ITU). *Methodologies for the Subjective Assessment of the Quality of Television Images*; Document Rec. ITU-R BT.500-14, 10/2019; ITU: Geneva, Switzerland, 2020.

54. International Telecommunication Union (ITU). *Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment*; Document Rec. ITU-T P.913, 03/2016; ITU: Geneva, Switzerland, 2016.
55. Huynh-Thu, Q.; Garcia, M.; Speranza, F.; Corriveau, P.J.; Raake, A. Study of Rating Scales for Subjective Quality Assessment of High-Definition Video. *IEEE Trans. Broadcast.* **2011**, *57*, 1–14. [[CrossRef](#)]
56. Wu, D.; Yuan, F.; Cheng, E. Underwater No-Reference Image Quality Assessment for Display Module of ROV. *Sci. Program.* **2020**, *2020*, 1–15.
57. Schober, P.; Boer, C.; Schwarte, L. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth. Analg.* **2018**, *126*, 1763–1768. [[PubMed](#)]
58. Okarma, K. Quality Assessment of Images with Multiple Distortions using Combined Metrics. *Elektron. Elektrotehnika* **2014**, *20*, 128–131. [[CrossRef](#)]
59. Ieremeiev, O.; Lukin, V.; Okarma, K.; Egiazarian, K. Full-Reference Quality Metric Based on Neural Network to Assess the Visual Quality of Remote Sensing Images. *Remote Sens.* **2020**, *12*, 2349. [[CrossRef](#)]
60. Zhai, G.; Min, X. Perceptual Image Quality Assessment: A survey. *Sci. China Inf. Sci.* **2020**, *63*, 211301. [[CrossRef](#)]

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).