# Investigating the Relationship between Image Quality and Crowdsourcing for Labeling

**LINA E. BUDDE[1], DAVID COLLMAR [2], UWE SÖRGEL[2] & DOROTA IWASZCZUK[1]**

*Abstract: The quality of training data and the performance of machine learning approaches using those data stand in direct correlation: Even the best algorithms are not able to compensate low-quality training data, making the use of high-quality training data essential. However, generating high-quality training data is not trivial, especially as these data are required in large quantities. A common approach is the use of paid crowdsourcing, where the resulting label quality is contingent upon two key factors: the performance of the workers and the intrinsic quality of the initial dataset utilized for these tasks. Limiting factors, such as poor radiometry that leads to low-quality images, must be taken into consideration when evaluating data acquired through crowdsourcing. We examine the relationship between image quality and the output performance of crowdsourcing tasks in the context of tree crown detection.*

## 1 Introduction

Due to the lack of large databases with reference data in remote sensing, manual annotation for the specific machine learning tasks is common (SCHMITT et al. 2023). Human-based annotations, which are the standard method, are time consuming and cost inefficient, especially when annotations are acquired by experts. Therefore, crowdsourcing can be used, which can provide an alternative at low cost (HIRTH et al. 2011) in a fast and easy manner (SARALIOGLU & GUNGOR 2020). While coming with these advantages, using paid crowdsourcing can also lead to quality issues due to various factors (ZHANG et al. 2016). The identification of those acquisitions of low quality is typically done via reference data, or, if not available, can be solved via "quality improvement after data collection" (ZHANG et al. 2016), where the same dataset is processed by different crowdworkers in order to collect redundant acquisitions. For the case of classifications, however, ambiguous class definitions or fuzzy borders might also influence label quality. The same issue might be present for the collection of outline geometries.

When using remote sensing imagery for such an acquisition task, where outline geometries are to be collected, the quality of the image data used for the annotations, i.e., the background imagery, influences the performance of crowdworkers (HASHMI et al. 2023). However, pre-processing steps such as ortho-image rectification or radiometric adjustments due to light conditions are common in remote sensing to reduce specific image effects or transform into geometric accurate images. As a drawback, this can result in both blur and noise in images (CHEN et al. 2020), resulting in many potential sources for low-quality areas. The identification of the latter and thereby the quality of imagery can reliably, although subjectively, be assessed by humans. However, objective machine-based evaluations of the

[1] Technische Universität Darmstadt, Fachbereich Bau- und Umweltingenieurwissenschaften, Fachgebiet Fernerkundung und Bildanalyse, Franziska-Braun-Str. 7, D-64287 Darmstadt, E-Mail: [lina.budde, dorota.iwazczuk]@tu-darmstadt.de

[2] Universität Stuttgart, Institut für Photogrammetrie, Geschwister-Scholl-Str. 24D, D- 70174 Stuttgart, E-Mail: [david.collmar, uwe.soergel]@ifp.uni-stuttgart.de

 DOI: 10.24407/KXP:1885661711

image quality remain a challenging task (YANG et al. 2023). Due to lack of error-free reference images, so called no-reference or blind image quality assessment metrics are preferred (RUBEL et al. 2022; YANG et al. 2023). However, some of these no-reference metrics still need additional information, e.g. about the respective type of distortion (YANG et al. 2023). For the application of these metrics in the field of remote sensing, IEREMEIEV et al. (2020) tested and combined metrics that require error-free references and are therefore impractical for real-world situations. In contrast, YAN et al. (2019) developed a specific gradient-weighted metric based on natural scene statistics and thus combine statistical, structural and contrast related image characteristics. RUBEL et al. (2022) proposed an approach to combine multiple no-reference metrics for the remote sensing domain. However, these are not tested within a realistic scenario. So, either natural image datasets (IEREMEIEV et al. 2020; RUBEL et al. 2022) or synthetic global image degradation (YAN et al. 2019) are used and compared with subjective scores.

In this paper, we want to investigate the relationship between label quality, measured via crowdsourcing, and image quality, measured via no-reference image quality metrics. In order to achieve realistic results, we perform both quality assessment processes with a real-world dataset, focusing on the application of simple and easy to use metrics. Furthermore, we define quality metrics derived from the results acquired via crowdsourcing on the same dataset, allowing for a direct comparison between all quality metrics. In summary, our contribution includes the development of a no-reference label quality metric and a workflow to identify causes for low quality labels with no-reference image quality metrics.

The paper is structured as follows. First, our workflow including all used methods is described in section 2. Section 3 contains the description of the used data. The results are presented in section 4. Finally, the conclusion can be found in section 5.

## 2 Methodology

An overview of our pipeline is presented in Figure 1. First, the image quality and the annotation quality are evaluated separately. Based on the resulting quality ranking of the images, the correlation of both, the image and annotation quality are analyzed. The specific quality assessment is described in the next sections.

### 2.1 Image Quality

Due to missing reliable reference information about the image quality, we focus on no-reference quality metrics. However, this metrics are often specialized for a certain image distortion type. Thus, we combined multiple of such metrics by calculating the mean value to consider different image effects. To validate the image quality metrics, they are applied on both, the raw images and the ortho-images. In both cases, a sliding window of a size of $15 \times 15$ pixels is used to combine the pixel-wise values to an image-wise score. To ensure comparability between different resulting single scores from the different metrics, they are scaled using min-max scaling based on the ortho-image results, where a low score means a high-quality image and a score near 1 represents an image with low quality. *1– score* is applied to the scores of MANIQA, edge intensity and local contrast after scaling so that they can be interpreted in the same way. In the following, the used metrics are shortly described.

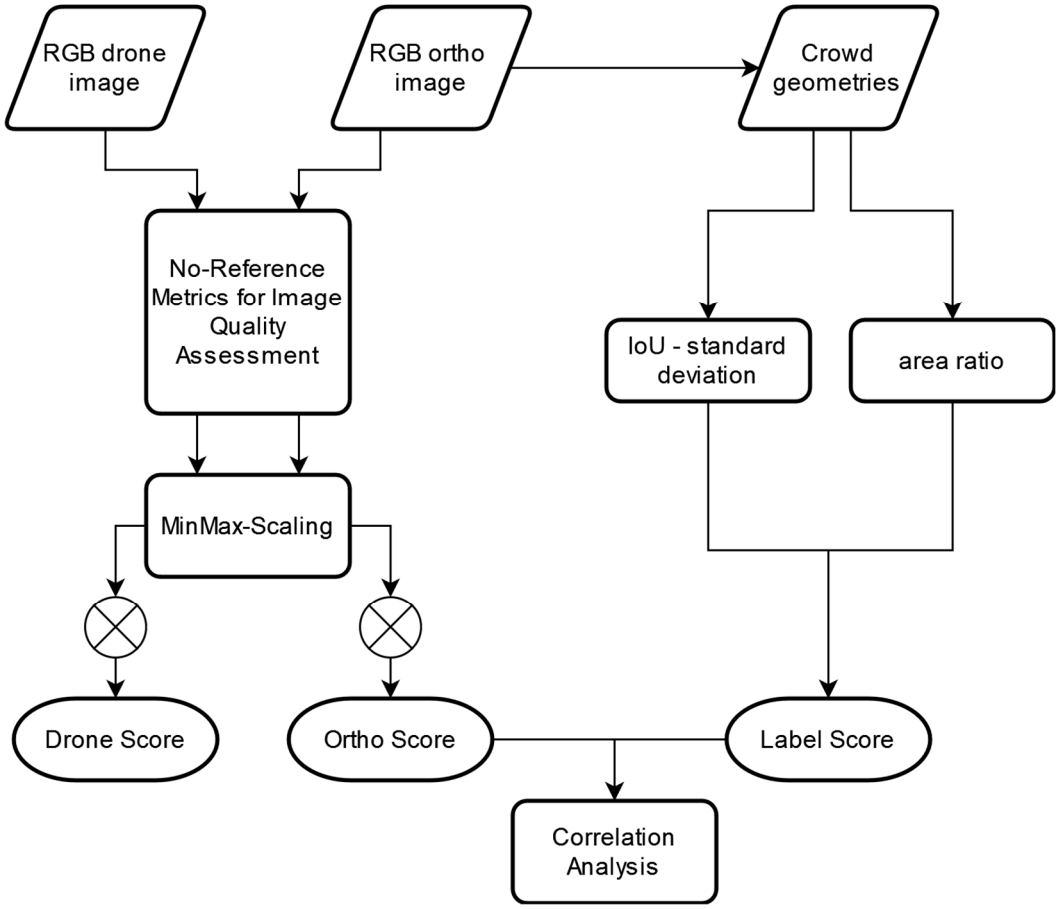


Fig. 1: Evaluation pipeline for the image and label quality. For the image quality, the raw drone images and the processed ortho-images are evaluated. After scaling, the different no-reference image metrics are averaged to determine the drone and ortho score, respectively. For the label quality, the standard deviation of the intersection over union is determined using a reference label and the crowd geometries. In addition, the no-reference metric, called area ratio, is calculated only based on the crowd geometries.

### 2.1.1 Edge Intensity

To consider the occurrence of edges, a Sobel filter is used for each channel and the mean value of the three channels is calculated (LI et al. 2017). To generate an image-wise edge score, the mean edge intensity is used within a sliding window. More clear edges in the image results in a higher edge intensity. However, edges significantly contribute to the interpretability of an image. Thus, a better image is associated with higher edge intensity.

### 2.1.2 Signal-to-noise Ratio and Local Contrast

Whereas, the contrast of an image can be determined by the normalized difference of the maximum and minimum gray value within a sliding window, the estimate of the signal-to-noise-ratio (SNR) based on the variance (XIA & CHEN 2015). Like the edge intensity, the values are calculated channel-wise and the mean value is used for the image score. For each sliding window, a channel-wise SNR and contrast are calculated by

$$contrast_c = \frac{\max(\mathcal{D}_c) - \min(\mathcal{D}_c)}{\max(\mathcal{D}_c) + \min(\mathcal{D}_c)} \quad (1)$$

$$SNR_c = \frac{\overline{D_c}}{\sqrt{1/(N-1)\ \sum(D_c - \overline{D_c})^2}} \quad (2)$$

where $D_c$ represents the channel-wise gray values and $N$ the number of pixels in the sliding window. In general, the SNR is applied in homogenous areas to determine the image noise. However, applying SNR on each sliding window, low SNR values can indicate inhomogeneous areas with high pixel variations due to e.g. noise or rich image content. Note that saturation effects can also lead to homogenous areas.

### 2.1.3 Blur Effect

Especially due to drone movements and ortho-image processing, blurring often occur. Thus, the metric introduced by (CRETE et al. 2007) can be used to determine the strength of the blurring. The effect of a low-pass filter is used to quantify the image blur. In image regions which are already blurry, the effect of low-pass filtering can be expected to be quite low. In contrast, the differences between the neighboring pixel values change significantly after applying low-pass filter in sharp image regions. With this approach the blur score is in the range [0, 1].

### 2.1.4 Shadow

To consider different light conditions the normalized saturation-value difference index (NSVDI) index can be used (MA et al. 2008). This index based on the saturation and value after color transformation from RGB to HSV. According to MA et al. (2008), the saturation $s$ and value $v$ is used

$$NSVDI = \frac{s - v}{s + v} \quad (3)$$

where the value range for this index is specified as [-1, 1].

### 2.1.5 BRISQUE

Compared to the previous metrics, the blind/referenceless image spatial quality evaluator (BRISQUE) score, developed by MITTAL et al. (2012), is based on a trained model. However, the trained parameters are public available to use BRISQUE without own training. The metric is based on locally normalized luminances to quantify naturalness of an image as well as the quality of distorted images. Therefore, the local mean intensity is subtracted and then normalized by a Gaussian weighted standard deviation (MITTAL et al. 2012). From this luminances, 36 features are calculated and mapped with a support vector regressor to a subjective quality score (MITTAL et al. 2012). The minimum value of BRISQUE is 0 which represents a higher quality image. A maximum value is usually given as 100 (DIEREND & GÖRNER 2022).

### 2.1.6 NIQE

The natural image quality evaluator (NIQE) score belongs to the same category of scores as BRISQUE (MITTAL et al. 2013). Thus, also NIQE use a natural scene statistic model based on locally normalized luminances. In contrast to BRISQUE, NIQE does not use distorted images and subjective scores for training (MITTAL et al. 2013). As for BRISQUE a value near zero corresponds to a better image, but an upper bound is not defined.

### 2.1.7 MANIQA

To consider a deep learning-based approach for quality assessment, the multi-dimension attention network for no-reference image quality assessment (MANIQA) is tested (YANG et al. 2022). Developed for the evaluation of GAN-based image degradation, this score considers the local and global interactions in channel and spatial dimensions (YANG et al. 2022). For the BRISQUE, NIQE and MANIQA scores the implementation from CHEN & MO (2022) is used. In contrast to BRISQUE and NIQE, CHEN & MO (2022) state that higher values are associated with a better image.

## 2.2 Annotation Quality

Similar to image quality, annotation quality needs to be measured via different metrics. Since "quality improvement after data collection" (ZHANG et al. 2016) is performed, redundant acquisitions are collected, allowing for an inherent quality assessment by quantifying the deviation in quality of acquisitions. For this, we make use of the principle of "wisdom of the crowd" (JIN et al. 2020), which states that, given constraints such as a heterogeneous crowd and a solvable task, the crowd as a whole can perform better than single experts (CHANDLER et al. 2013). Based on this principle, and the fact, that the crowd acquisition is performed via polygon outlines, we define two quality metrics.

### 2.2.1 IOU-STD

Firstly, for the case where reference data are existing and might be used, acquisition quality can be calculated via comparison to these reference data. Given a dataset consisting of $k$ image sections with $n$ redundant acquisitions per section, we calculate the intersection over union (from now on referred to as $IoU$) between all $n$ acquisitions and their respective references, leading to $n$ IoU values. Subsequently, we calculate the standard deviation of all IoU values per image section, leading to $k$ values, where each value describes the overall deviation in crowd acquisitions per image section. This approach might be sensitive for small sample sizes, however, choosing $n$ big enough allows to reduce the impact of outliers, following Condorcet's theorem (CONDORCET 1785). Low values indicate low deviations among all crowd acquisitions, signifying results of high similarity, and thereby making low scores favorable.

### 2.2.2 AREA-RATIO

Secondly, we also highlight the case if ground truth data are not available. Since the calculation of the previously explained IoU deviation requires the existence of reference data, a similar approach is not applicable here. However, reference data are not necessarily needed for a deviation assessment. Instead, simple parameter such as surface area or perimeter of acquisitions can be calculated and used for a quantification of deviation in acquisitions. However, such parameters do not consider the actual shape of the acquisitions, which might lead to imprecise or superficial results. Instead, a polygon integration as described in COLLMAR et al. (2023) is performed on all acquisitions. We integrate all acquisitions twice, using a too high ($t_h$) and too low ($t_l$) threshold for the integration, deliberately not picking the optimal threshold. This leads to two output shapes: One integrated shape that appears too large ($A_{t_l}$), and another one, that appears too small ($A_{t_h}$). However, due to the nature of the integration, which is performed via binary voting, the resulting shapes are representative of both a small part of crowdworkers (that is $n - t_h$) and a large part of crowdworkers (that is $n -$

$t_l$), leading to two different shapes that allow a direct comparison. These two shapes are then used for calculating the ratio between each other, assigning a single value between 0 and 1 to all $k$ image sections:

$$R_{Area} = \frac{A_{t_h}}{A_{t_l}}, \text{with } t_h > t_l \quad (4)$$

A value of 1 indicates a perfect match between both shapes, describing no deviation in crowd acquisitions. A value close to 0 on the other hand can be attributed to large discrepancies in both shapes, indicating large variations in the acquired geometries. Figure 2 visualizes both the choice of a too small and too large threshold next to the optimal one, showing the resulting variance in surface area.

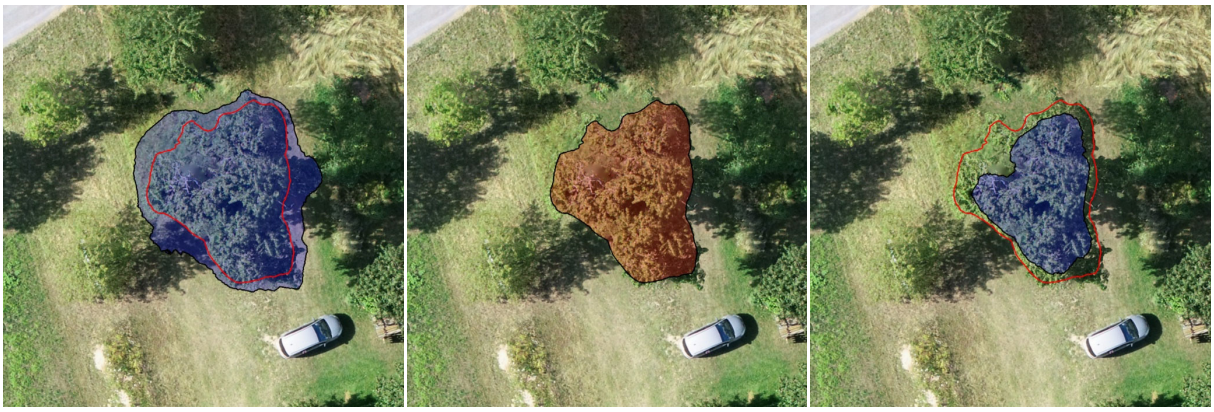

Fig. 2: From left to right: Output polygon after integration with a threshold set too low ($A_{t_l}$), integration using an optimal threshold, and integration with a threshold set too high ($A_{t_h}$). The optimal threshold result is indicated in both the low and high threshold images for direct comparison.

### 2.3 Evaluation

To validate the image quality scores, the selected image quality metrics can be applied before and after pre-processing of the images. Thus, Welch's t-Test is used to check for significant change due to the pre-processing. For the pre-processed images, the Pearson correlation coefficients *Corr* are calculated between the different image quality scores and the label scores as well as between the mean image quality score and the label scores based on the covariance matrix *cov*.

$$Corr = \frac{cov_{ij}}{\sqrt{cov_{ii}\, cov_{jj}}} \quad (4)$$

## 3 Data

### 3.1 Imagery

The dataset comprises 97 RGB image pairs, each consisting of raw imagery and corresponding sections from the processed orthomosaic of the same area. Each image section features a single tree and its surrounding area. All raw images were captured by a DJI FC6310R camera in combination with a DJI Phantom 4 RTK. The flight was conducted in an orchard area in southern Germany and the flight height was set to 100 m, thereby resulting in a GSD of around 2.5 cm.

### 3.2 Crowd acquisition

All 97 ortho-image sections were presented to 100 crowd workers each via microWorkers.com, a platform that manages the recruitment and payment of workers for crowdsourcing tasks (HIRTH et al. 2011). Workers were tasked with collecting the outline geometries of trees, receiving a payment of \$0.15 for the processing of 5 image sections. This resulted in total costs of $100 \text{ acquisitions} \cdot 97 \text{ sections} \cdot \$0.03 = \$291$ for 9,700 outline polygons. Figure 3 illustrates such an image pair, as well as the corresponding annotations.

Fig. 3: From left to right: raw drone image, ortho-image and the geometries collected by crowdworkers on the ortho-image. Mainly caused by the shadow, at the right side of the tree, blurring occurs due to the ortho processing. Thus, it is harder to distinguish between the ground and the tree resulting in a higher variation of the annotations in this part of the tree.

## 4 Results

This section contains the results of the individual evaluations as well as of the correlation between the image and label quality.

### 4.1 Image Quality Metrics

As can already be seen in Figure 3, blurred areas can occur after ortho processing. This effect of image degradation can also be observed in other images of the dataset. Figure 4 shows the results of the final scores for the raw images and the processed ortho-images. The scores for the raw images are lower than the scores of the ortho-images. Thus, the different processing levels can be differentiated by the selected scores. This can be also validated by the t-Test result and the histogram. The probability of a common mean value is $4 \cdot 10^{-38}$. Thus, the distributions of the mean image quality score for the raw drone images and the processed ortho-images are different.

### 4.2 Annotation Quality

Both quality measures, i.e., the standard deviation of the IoU values in comparison to the reference, as well as the previously defined area ratio, were calculated for all 97 image sections. Since low values are preferable for the IoU standard deviations, but high values are favored for the area ratio, a correlation analysis between the two was conducted. This calculation led to an overall correlation of -0.54 between the two methods, confirming the expected, negative correlation. This correlation is further visualized in Figure 5, which plots both metrics against

each other. For clearer visualization, results were sorted and smoothed, thereby indicating a higher correlation than the one actually achieved of around -0.54.

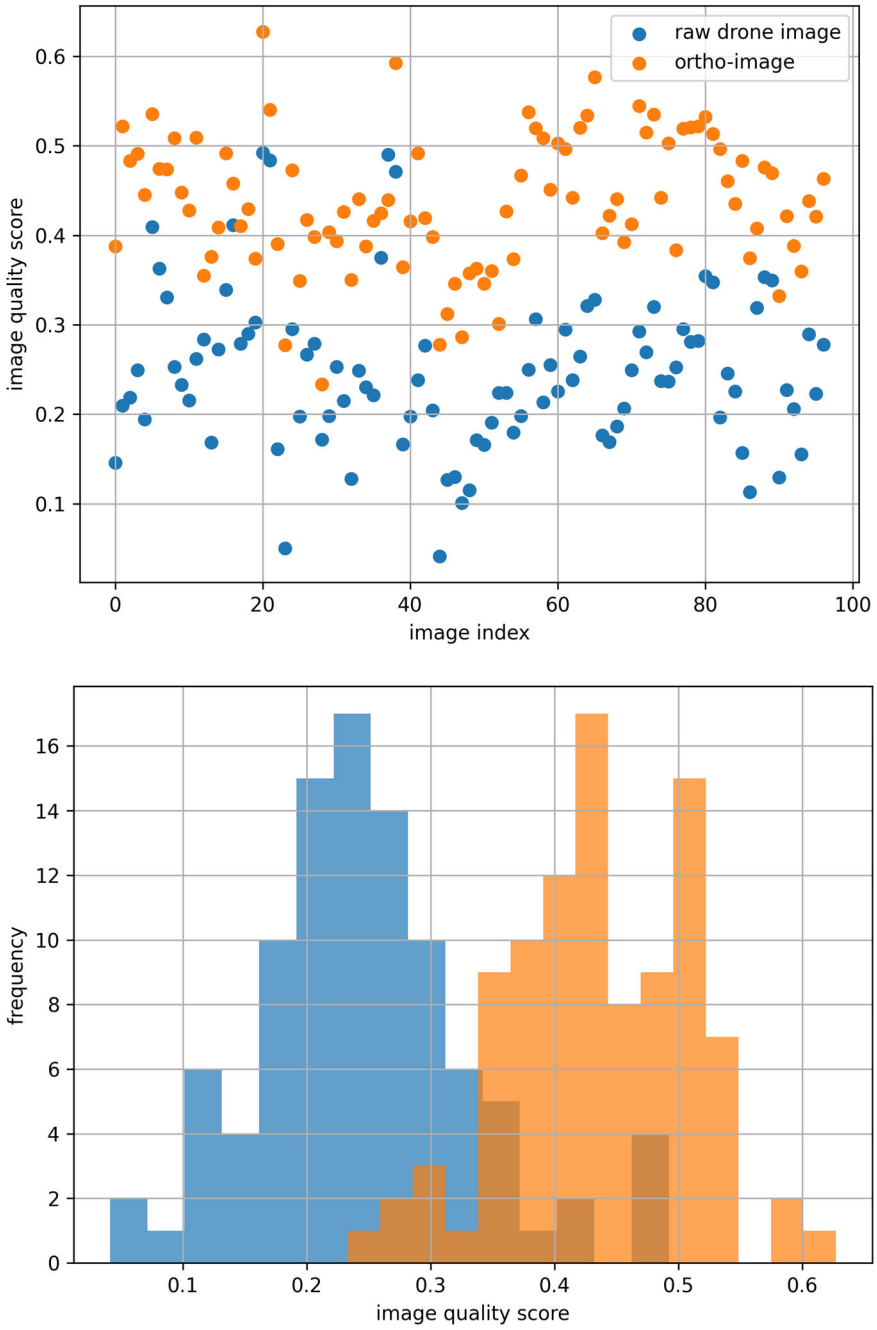


Fig. 4: Top: Scores of the individual images. Bottom: Distribution of the raw and ortho-image scores.

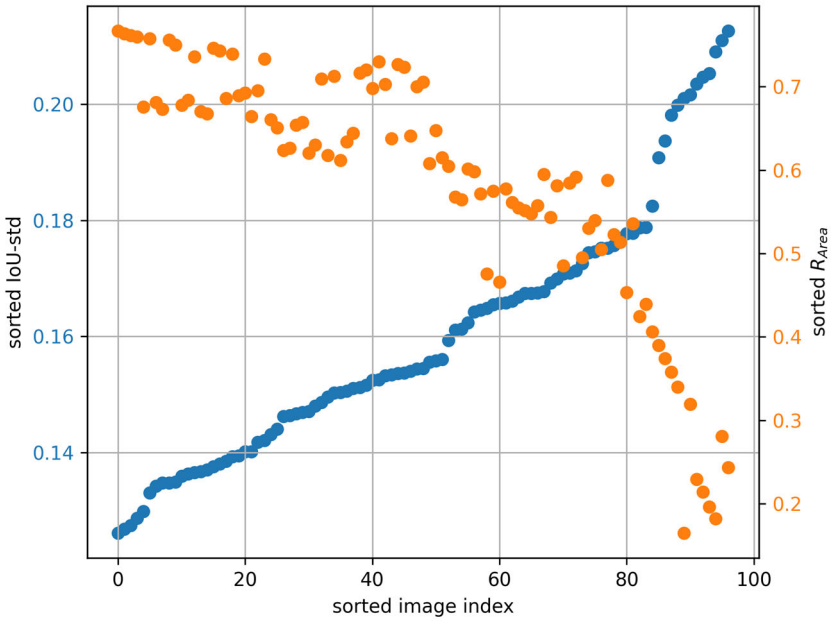


Fig. 5: Comparison between the IoU-std and the $R_{Area}$ sorted by the increased IoU-std values.

Figure 5 also indicates the tendency of both metrics to be more sensitive for image sections with bad acquisitions: Both the standard deviation of IoU as well as sorted area ratio show a steeper gradient for larger indices, i.e., those of high deviation and low area ratio. This observation suggests that both metrics penalize poor-quality acquisitions effectively, which is a desirable outcome, helping to identify image sections that posed most problems to crowdworkers.

### 4.3 Correlation Analysis

We analyzed the correlation between the mean image quality score and the label quality scores, namely IoU-std and area ratio, as well as the correlation between the individual image quality scores and again the two label scores. With the comparison with the individual scores, we are able to analyze which metric and thus which distortions may influence the consistence of the tree labels.

The resulting correlation coefficients can be found in Table 1. Based on the evaluation with the IoU-std, the absolute highest correlation is reached by the edge intensity score followed by the contrast and BRISQUE. The scores from the other metrics, in contrast, do not have a correlation with the label quality. However, the negative sign of the correlation implies that the quality of the crowd labels increase if the images have a decreased strength of edges. The example shown in the upper row of Figure 6 displays a tree with high contrast and distinct edges. The image quality score of 0.23 indicates an image of rather good quality. Nevertheless, the IoU-std score shows a larger deviation of the crowdworker labels to the reference label, whereby outliers have already been filtered out.

This relationship can be also observed in the second example in the bottom of Figure 6. There, the tree barely stands out from the background and the edge intensity is low. However, the consistence of the labels is still good. When looking at the edge images, it is noticeable that the edges in the upper example also occur in the immediate vicinity of the targeted tree, which can make labeling more difficult. In contrast, the area surrounding the tree in the example below is relatively free of edges.

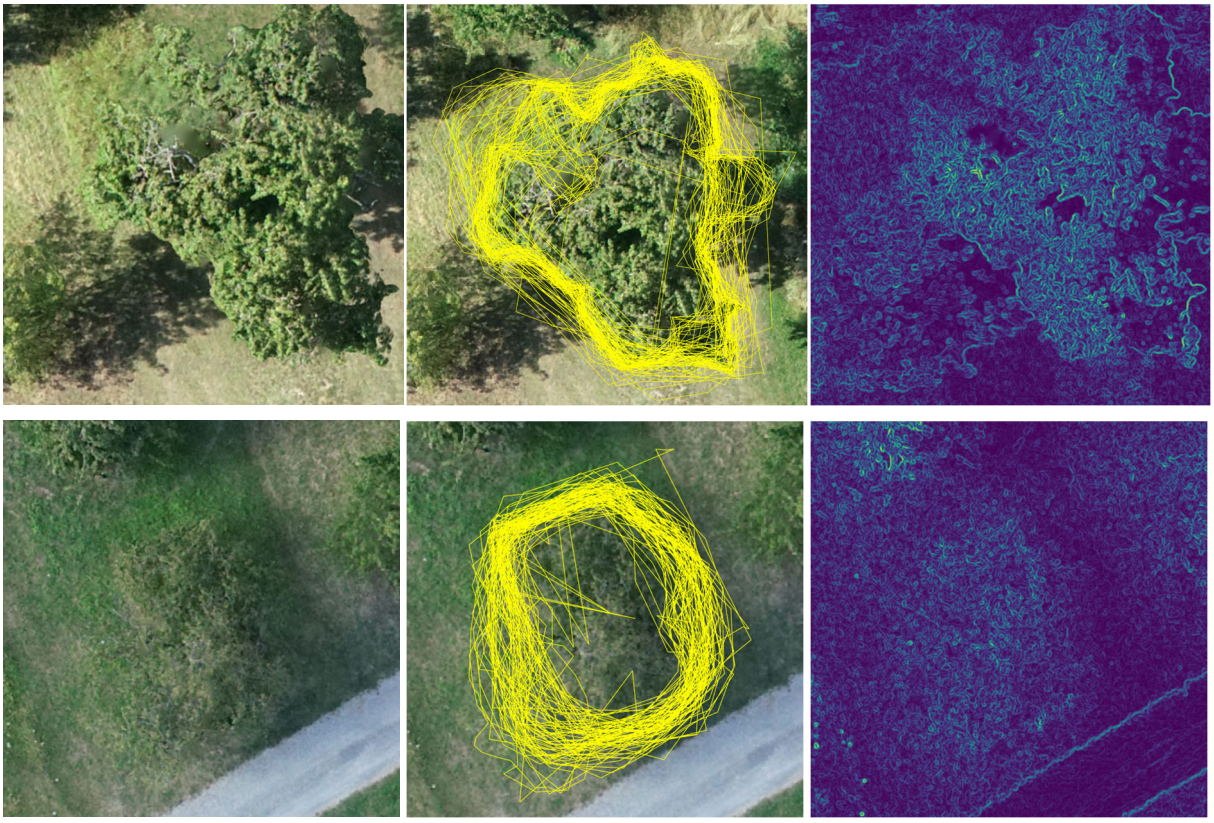

Fig. 6: From left to right: Ortho-image, crowd annotations with removed outliers, edges from Sobel filter.

A similar correlation can be observed between the image quality scores and the area ratio score. To ensure that in both cases, the area ratio score and the image quality scores are interpreted in a way that lower values indicate better quality, Table 1 uses $1 - R_{Area}$ for the comparison of correlation coefficients. Most image quality scores appear to correlate in a similar fashion with both label quality parameters, validating the novel no-reference approach, i.e., the area ratio. However, some differences are notable: the edge intensity, BRISQUE and contrast score correlate lower compared to IoU-std. More interestingly, the MANIQA score has a moderate positive correlation with the $1 - R_{Area}$ score. This results in a cancellation of the single correlations, so that overall no correlation with the averaged image quality values can be achieved. This highlights the complexity of the relationships and the difficulty of finding objective metrics for quality evaluation.

## 5 Conclusion

To the best of our knowledge, this is the first evaluation of the relationship between no-reference image quality metrics and the quality of crowdsourced annotations for a remote sensing task, incorporating a novel no-reference annotation quality metric. With the proposed area ratio metric for the label quality evaluation, we overcome the limitations of the standard deviation of the IoU which depends on the availability of a reference label and its already compromised quality. Furthermore, our findings indicate interrelationships between metrics derived from image quality and those derived from annotation quality. The ensuing correlation analysis provides further insights into the precise nature of these relationships, such as the

identified negative correlation between edge intensity and annotation quality metrics. Given that our analysis considered the entire image for the calculation of image quality metrics, local variations were not accounted for, which simplifies the interpretation due to the complexity of various image quality scores. Nonetheless, future studies might benefit from a more detailed examination, which could prove valuable in making more generalized statements.

Tab. 1: Correlation coefficients between one of the listed image quality scores and the label quality scores

| Image Quality Score | IoU-std | $1 - R_{Area}$ |
|---|---|---|
| Mean | -0.31 | -0.05 |
| BRISQUE | -0.28 | -0.11 |
| NIQE | 0.04 | 0.03 |
| MANIQA | 0.01 | 0.38 |
| SNR | -0.07 | 0.08 |
| Blur | 0.06 | 0.02 |
| Edge Intensity | -0.41 | -0.30 |
| Contrast | -0.31 | -0.18 |
| Shadow | 0.07 | 0.03 |

## 6 References


CHANDLER, J., PAOLACCI, G. & MUELLER, P., 2013: Risks and rewards of crowdsourcing marketplaces. Handbook of Human Computation, Michelucci, P. (eds), Springer, New York, NY. https://doi.org/10.1007/978-1-4614-8806-4_30.

CHEN, C. & MO, J., 2022: IQA-PyTorch: Pytorch toolbox for image quality assessment. https://github.com/chaofengc/IQA-PyTorch, last access 22.01.2024.

CHEN, G., PEI, Q. & KAMRUZZAMAN, M.M., 2020: Remote sensing image quality evaluation based on deep support value learning networks. **83**(115783). https://doi.org/10.1016/j.image.2020.115783.

COLLMAR, D., WALTER, V., KÖLLE, M. & SÖRGEL, U., 2023: From multiple polygons to single geometry: Optimization of polygon integration for crowdsourced data. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., **X**-1/W1-2023, 159-166, https://doi.org/10.5194/isprs-annals-X-1-W1-2023-159-2023.

CONDORCET, M.D., 1785: Essai sur l'application de l'analyse a la probability des decisions rendues a la plurality des voix. paris: De l'imprimerie royale. translated in 1976 to "essay on the application of mathematics to the theory of decision-making". Selected Writings, Baker, KM (eds), Indianapolis, IN, Bobbs Merrill.

CRETE, F., DOLMIERE, T., LADRET, P. & NICOLAS, M., 2007: The blur effect: perception and estimation with a new no-reference perceptual blur metric. Human Vision and Electronic Imaging SPIE Proceedings, Rogowitz, B.E., Pappas, T.N., Daly, S.J. (eds.), **XII**(64920I), https://doi.org/10.1117/12.702790.

DIEREND, H. & GÖRNER, H., 2022: Data quality metrics. https://quality.nfdi4ing.de/en/latest/image_quality/BRISQUE.html, last access 23.01.2024.

HASHMI, A.A., AGAFONOV, A., ZHUMABAYEVA, A., YAQUB, M. & TAKÁČ, M., 2023: In quest of ground truth: Learning confident models and estimating uncertainty in the presence of annotator noise. http://arxiv.org/pdf/2301.00524v1.

HIRTH, M., HOßFELD, T. & TRAN-GIA, P., 2011: Anatomy of a crowdsourcing platform-using the example of microworkers.com. 2011 Fifth IEEE international conference on innovative mobile and internet services in ubiquitous computing, Seoul, Korea (South), 322-329, https://doi.org/10.1109/IMIS.2011.89.

IEREMEIEV, O., LUKIN, V., OKARMA, K. & EGIAZARIAN, K., 2020: Full-reference quality metric based on neural network to assess the visual quality of remote sensing images. Remote Sensing, **12**(15), 2349, https://doi.org/10.3390/rs12152349.

JIN, Y., CARMAN, M., ZHU, Y. & XIANG, Y., 2020: A technical survey on statistical modelling and design methods for crowdsourcing quality control. Artificial Intelligence, **287**(103351), https://doi.org/10.1016/j.artint.2020.103351.

LI, S., YANG, Z. & LI, H., 2017: Statistical evaluation of no-reference image quality assessment metrics for remote sensing images. ISPRS Int. J. Geo-Inf., **6**(5), 133, https://doi.org/10.3390/ijgi6050133.

MA, H., QIN, Q. & SHEN, X., 2008: Shadow segmentation and compensation in high resolution satellite images. IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium, **2**, II–1036–II–1039, https://doi.org/10.1109/IGARSS.2008.

MITTAL, A., SOUNDARARAJAN, R. & BOVIK, A.C., 2013: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters, **20**(3), 209-212, https://doi.org/10.1109/LSP.2012.2227726.

MITTAL, A., MOORTHY, A.K. & BOVIK, A.C., 2012: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing, **21**(12), 4695-4708, https://doi.org/10.1109/TIP.2012.2214050.

RUBEL, A., IEREMEIEV, O., LUKIN, V., FASTOWICZ, J. & OKARMA, K., 2022: Combined no-reference image quality metrics for visual quality assessment optimized for remote sensing images. Applied Sciences, **12**(4), 1986, https://doi.org/10.3390/app12041986.

SARALIOGLU, E. & GUNGOR, O., 2020: Crowdsourcing in remote sensing: A review of applications and future directions. IEEE Geoscience and Remote Sensing Magazine, **8**(4), 89-110, https://doi.org/10.1109/MGRS.2020.2975132.

SCHMITT, M., AHMADI, S.A., XU, Y., TAŞKIN, G., VERMA, U., SICA, F. & HÄNSCH, R., 2023: There are no data like more data: Datasets for deep learning in earth observation. IEEE Geosci. Remote Sens. Mag., **11**(3), 63-97, https://doi.org/10.1109/MGRS.2023.3293459.

XIA, Y. & CHEN, Z., 2015: Quality assessment for remote sensing images: Approaches and applications. 2015 IEEE International Conference on Systems, Man, and Cybernetics, 1029-1034, https://doi.org/10.1109/SMC.2015.186.

YAN, J., BAI, X., XIAO, Y., ZHANG, Y. & LV, X., 2019: No-reference remote sensing image quality assessment based on gradient-weighted natural scene statistics in spatial domain. J. Electron. Imag., **28**(01), 1, https://doi.org/10.1117/1.JEI.28.1.013033.

YANG, P., STURTZ, J. & QINGGE, L., 2023: Progress in blind image quality assessment: A brief review. Mathematics, **11**(12), 2766, https://doi.org/10.3390/math11122766.

YANG, S., WU, T., SHI, S., LAO, S., GONG, Y., CAO, M., WANG, J. & YANG, Y., 2022: Maniqa: Multi-dimension attention network for no-reference image quality assessment. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1190-1199, https://doi.org/10.1109/CVPRW56347.2022.00126.

ZHANG, J., WU, X. & SHENG, V.S., 2016: Learning from crowdsourced labeled data: a survey. Artificial Intelligence Review, **46**(4), 543-576, https://doi.org/10.1007/s10462-016-9491-9.