Enhancing Data Quality in Crowdsourcing for Tree Outline Acquisition in Aerial Imagery via CNN-Based Real-Time Feedback

David Collmar[1], Volker Walter[1], Uwe Sörgel[1]

[1] Institute for Photogrammetry and Geoinformatics (ifp), University of Stuttgart, Germany
(david.collmar, volker.walter, uwe.soergel)@ifp.uni-stuttgart.de

Keywords: Crowdsourcing, Data Enhancment, Cost Optimization, CNN-Based Quality Control.

Abstract

We propose a method to improve data quality in paid crowdsourcing by leveraging CNN-based real-time feedback. Data acquired through paid crowdsourcing often suffers from inconsistencies or inaccuracies as workers prioritize task completion speed over precision to maximize earnings. To address this issue, we developed a lightweight, two-branch CNN that evaluates and provides quality feedback on polygon acquisitions of tree outlines in aerial imagery. As workers modify their polygons, the CNN predicts a quality score, displayed as a traffic light signal (red, yellow, green), indicating whether adjustments are needed. Our study compares a test group receiving this feedback with a control group without feedback. Results show that the test group achieves a notably higher average Intersection over Union (IoU) score as well as a lower standard deviation, indicating improved quality and consistency. By integrating the results of multiple workers, the test group achieves even better data quality with fewer samples than the control group. This approach reduces the need for redundant data acquisition, demonstrating its potential for time and cost savings in large-scale data collection campaigns.

1. Introduction

The manual acquisition of high-quality data, including geospatial applications, has traditionally been the domain of experts, which still holds true in an era of increasing applications of machine learning (Wang et al., 2023, Ostyakova et al., 2023). This desire for high-quality data includes the field of remote sensing (Wang et al., 2023): Although expert annotations ensure high quality, they remain labor- and cost-intensive, impeding scalability (Rasmussen et al., 2022, Ostyakova et al., 2023). Further problems may arise from systematic errors such as annotator bias or inconsistencies across multple acquisitions (Rasmussen et al., 2022). A scalable and cost-effective alternative, that cancels out effects like systematic errors, is paid crowdsourcing (Brabham, 2013), which is used in many fields such as natural language processing or computer vision (Zhang, 2022). Crowdsourcing in general distributes tasks to a large, diverse pool of individuals, who may be paid or act on a voluntary basis (Brabham, 2013). Voluntary crowdsourcing relies on intrinsic motivation (Hossain, 2012), i.e., interest in the project itself. Subsequently, voluntary crowdsourcing not only has the advantage of no or limited monetary costs, but the intrinsic motivation can lead to higher-quality output (Rogstadius et al., 2011). Generating such intrinsic motivation might be challenging, leading to paid crowdsourcing as a straightforward alternative. However, the financial incentive often leads to a mostly extrinsic motivation that can have a negative influence on output quality: Workers prioritize the speed of their task completion over the thoroughness of their submissions in order to maximize their financial gain (Chandler et al., 2013). This behavior can lead to inconsistent or deliberately incorrect data submissions, with up to 45% of workers producing low-quality output (Vuurens et al., 2011), making data quality a persistent issue in paid crowdsourcing (Kobayashi et al., 2022).

In order to prevent these low-quality submissions, different quality control mechanisms can be implemented to ensure data quality (Jin et al., 2020). These might range from simple qualification tests and hidden tests (short extra tasks as consistency checks) to sophisticated truth inference approaches as anti-spam measures (Cui et al., 2021). Qualification tests add to the workload of crowdworkers, while hidden tests raise the total cost, since more tasks need to be performed (Zheng et al., 2017). Furthermore, simple anti-spam measures such as these tests can be circumvented by malevolent crowdworkers (Zhu and Carterette, 2010). Worker modeling can help to identify these workers; however, worker modeling relies on redundant results of the same worker, i.e., observations over multiple completed tasks (Zheng et al., 2017). Since paid crowdsourcing typically relies on crowdsourcing platforms such as Microworkers.com (Hirth et al., 2011), where millions of workers are registered (Microworkers, 2024), approaches based on worker redundancy may be unfeasible, due to the assumed low likelihood of the same workers participating in multiple campaigns. Alternatively, approaches based on task redundancy instead of worker redundancy can be employed (Zhang et al., 2016), aiming to leverage the principle of ”wisdom of the crowd” (Jin et al., 2020). Here, the same task is performed by multiple workers, and results are subsequently integrated, e.g., via majority voting (Zheng et al., 2017). This integration can help to filter out spammers or mitigate their effects, thereby improving data quality (Zhang et al., 2016). For the acquisition of geometric outlines, a common task in remote sensing, research has shown that an increase in redundant datasets with subsequent integration leads to higher output quality (Collmar et al., 2023). Specifically, this study found that while quality continues to improve with high levels of redundancy in acquisitions, the benefits diminish relative to the rising costs (Collmar et al., 2023). Nevertheless, such high levels of redundancy are generally undesirable due to their additional costs in terms of both time and money. For both mentioned cases, i.e., worker or task redundancy, this creates the unwelcome trade-off between cost and output quality. Achieving higher quality necessitates high redundancy, driving up financial costs, whereas limiting redundancy to reduce expenses might compromise the output quality.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-197-2025

Another approach is to provide real-time feedback and peer evaluation, which research has shown can substantially improve task quality among crowdworkers (Dow et al., 2012). While peer evaluation enhances quality, its integration into crowdsourcing can be challenging as well as resulting in extended task durations and higher costs (Collmar et al., 2024). However, lightweight CNN models, capable of processing data rapidly and efficiently, support this need well. These models have been applied in a variety of applications like object detection (Chandana and Ramachandra, 2022) or emotion recognition (Ozdemir et al., 2019), where real-time feedback is necessary. In general, CNNs are widely used in computer vision and remote sensing for both real-time and non-real-time tasks, such as classification and segmentation (Liu et al., 2023). CNNs offer a practical solution for large-scale applications where manual work is unfeasible due to dataset size or cost (Liu et al., 2023, Chen et al., 2023).

In this paper, we aim to combine the strengths of both approaches, i.e., CNNs and manual acquisitions, by including a real-time CNN quality check to provide immediate feedback to crowdworkers. This approach aims to achieve high data quality while maintaining cost-effectiveness by minimizing the need for redundant acquisitions. The paper is structured as follows. First, we explain our methodology and introduce the datasets used, followed by an explanation of the training process and experimental setup. The results consist of two parts: change in data quality both with and without polygon integration, including a redundancy and cost analysis. We then discuss limitations and motivate future research, concluding with a summary of our findings.

2. Methodology

In order to combine the advantages of manual acquisition with the efficiency of automated processes, we propose a approach that leverages crowdsourcing and real-time CNN feedback for the example of acquisition of tree outlines from aerial images via polygons. While we focus on tree outlines in this setup, the method could be adapted for various types of geometric data.

Crowdworkers are asked to use a web-based interface to annotate tree crowns in aerial images by creating polygon geometries that capture tree outlines. To assess the quality of these polygons, a CNN evaluates the acquired polygons after every change performed by the user (i.e., adding or removing polygon vertices) by calculating a score that reflects the accuracy of the tree outline described by the polygon. For this, we use a two-branch CNN inspired by the solution proposed by Shi et al. (2017), and convert all crowd-acquired polygons to binary masks as input. Although they also propose a multi-scale network, we prefer their two-branch solution: while it performs only slightly worse in terms of quality, it addresses speed limitations necessary for a real-time response, offering an efficient trade-off. Furthermore, as our setup should allow for simultaneous communication between multiple crowdworkers and the server handling the CNN operations, computational resources may need to be divided among users. To manage this, we use a lightweight CNN model and optimize backend processes to ensure that each worker receives real-time feedback, enabling a scalable solution that maintains high performance.

Our adjusted network architecture therefore consists of two branches, each following a sequence of alternating Conv layers and MaxPool layers, arranged as Conv → MaxPool → Conv → MaxPool → Conv → MaxPool. This structure progressively reduces the spatial dimensions, allowing the network to retain critical features while minimizing computational load. Before being processed by the network, each input consists of the original aerial image and a corresponding binary mask of a polygon, which are separately fed into the two branches of the CNN. Given our input image resolution of 416x416, as will be explained in the next section, we found that this setup effectively captures the necessary features without requiring additional layers. After the branches are concatenated, a single dense layer is applied, enabling efficient feature consolidation without the need for additional fully connected layers, further improving real-time performance for parallel computation of multiple workers. Figure 1 provides a detailed overview of the adjusted architecture, including input shapes and layer configurations, highlighting our focus on optimizing both speed and memory usage for the proposed real-time application.

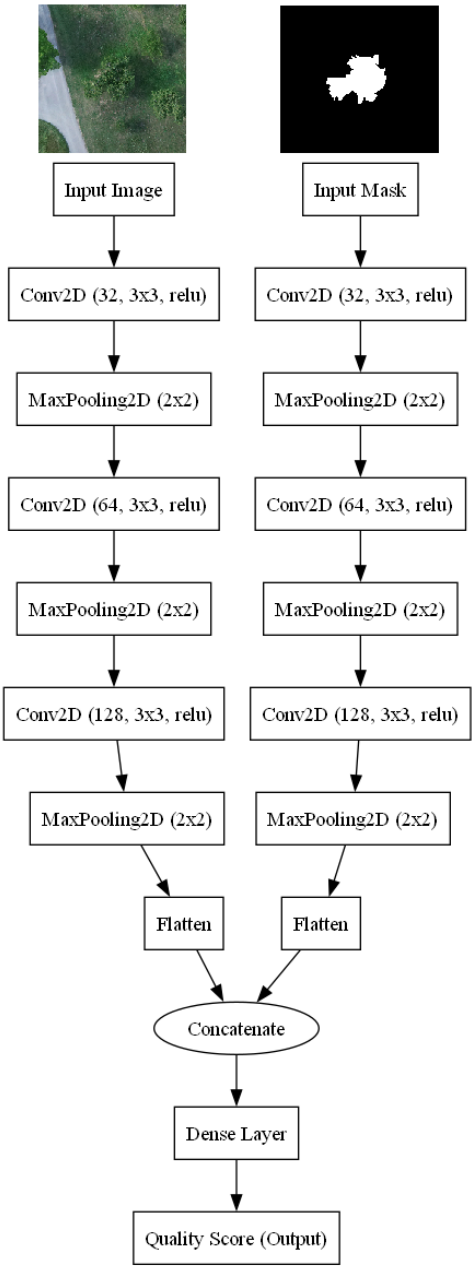

Figure 1. Proposed two-branch CNN architecture.

The output, i.e., a quality score between 0 and 1, can then not only be used to estimate the quality of the acquired polygon, but to indicate if an applied change (adding or removing a polygon vertex) yields benefits. To keep the feedback intuitive and to manage minor variations from the CNN, we use a simple ’traffic light’ system: red, yellow, or green lights signal to workers whether further refinement is required.

3. Dataset & Training

For this study, we used two distinct datasets, both consisting of image sections from a large-scale orthomosaic generated using drone imagery captured over different orchards in southern Germany. Each image section measures 416x416 pixels. Dataset A consists of 474 such image sections, whereas Dataset B only contains 50 image sections. Dataset A was used for training and validation of the CNN described above, while Dataset B was used for the actual study. Both datasets were captured under differing conditions: Dataset A was collected in the morning on a sunny day, resulting in visible shadows, whereas Dataset B was captured around noon on a cloudy day, producing smaller, less intense shadows and a generally different radiometry. Previous studies on manual tree crown annotation have shown that changes in image quality, resolution, and environmental conditions can influence the accuracy and consistency of manual annotations (Steier et al., 2024, Budde et al., 2024). Therefore, these variations in the two datasets are expected to enable an evaluation of our model’s robustness and adaptability, potentially offering insights into its generalizability across different conditions. Figure 2 shows two examples that highlight the aforementioned variations.

Figure 2. Examples from Dataset A (left) and Dataset B (right).

The training of the CNN was performed with Dataset A, with 80% of sections randomly assigned to training and the remaining 20% used for validation at the start of each epoch. A total of 15,218 manual polygon acquisitions of varying levels of detail from a previous crowdsourcing campaign were utilized as masks for training. For the quality parameter, we used the Intersection over Union (IoU) in respect to the available ground truth. During training, the IoU between the user-provided polygon and the ground truth mask is computed and used as the target, allowing the CNN to learn to predict a quality score that correlates with IoU rather than explicitly calculating it. A key advantage of predicting a quality score instead of explicitly computing IoU is that it enables the use of the full range of training data, including both high- and low-quality annotations. This ensures that the model learns to distinguish varying levels of annotation accuracy rather than relying solely on perfect examples, allowing it to generalize better across varying input qualities, including lower-quality acquisitions that would otherwise be difficult to assess. Since no ground truth is available during deployment, this generalization is essential for providing reliable quality feedback in real-world scenarios. The training process was executed over 10 epochs with a batch size of 16 for a good balance between the speed of training and memory efficiency. Figure 3 illustrates the training and validation loss over epochs 3 to 10 for the proposed CNN architecture. The plot begins at epoch 3 to exclude the larger fluctuations in the initial training epochs.

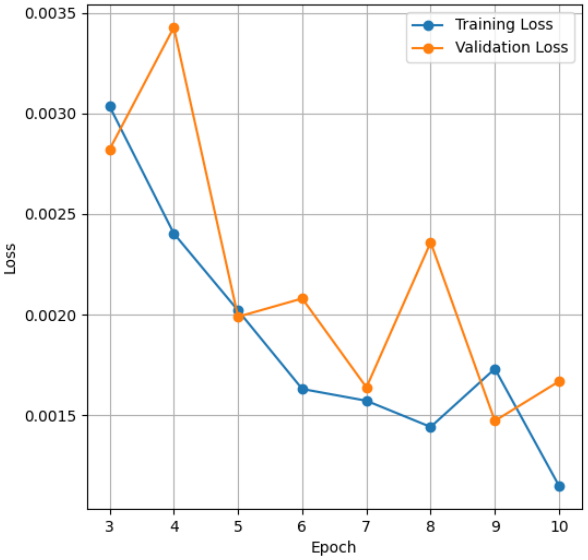

Figure 3. Training and validation loss for epochs 3 to 10.

Interestingly, although the training loss decreases further in epoch 10, subsequent epochs remain mostly stable. This stability, combined with the low final loss values, indicates that the model achieves a strong alignment between predicted and actual outputs. This demonstrates good performance within the training dataset, while maintaining a low risk of overfitting.

4. Experimental Setup

In order to evaluate the impact of our CNN on data quality in crowdsourcing, we conducted two separate crowdsourcing campaigns: one in a traditional way, without any CNN checks included, and one with CNN-based quality checks. For this, we established two distinct groups: a control group and a test group. The control group operated completely without CNN checks in order to assess user performance without any external input. In contrast, the test group operated with quality checks provided by the proposed CNN. This setup enables a direct comparison of task accuracy between the groups in order to measure the potential impact of the CNN-based quality checks on the worker.

4.1 Webtool

Both groups used a web tool that allows acquisitions of geometries via polygons by means of simple clicking to add or remove vertices. Simple consistency checks, such as setting the lower limit of polygon vertices to five, were implemented for both groups. The web tool for the test group was extended by the following feedback mechanism: With each modification of the polygon geometry, i.e., adding or removing a node, the current polygon is converted to a mask and sent to the backend. Then, a quality prediction using the two-branch CNN is performed on the server. The resulting quality estimation is then sent back to

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-197-2025 | © Author(s) 2025. CC BY 4.0 License.

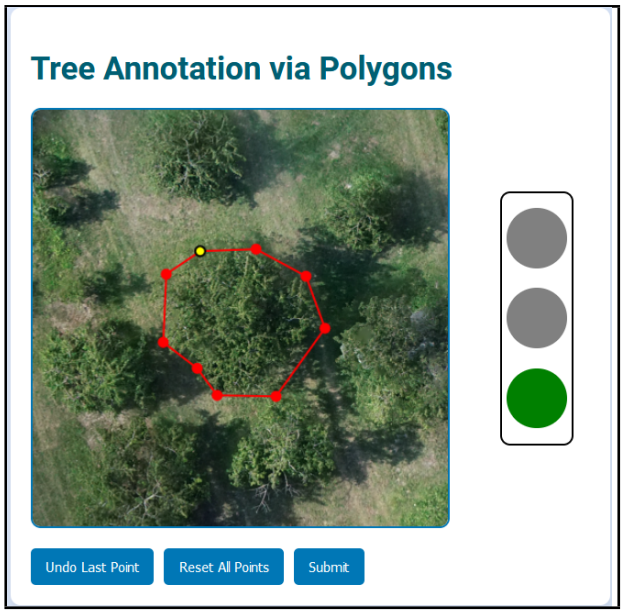

Figure 4. Web tool with example acquisition.

the client, where a traffic light indicator informs the user of the polygon’s quality, as shown in Figure 4.

The traffic light thresholds were chosen as follows. A quality score of 0.8 was set for a green light, as IoU scores around this value approach a saturation point, where further improvements add little value (Collmar et al., 2023). A threshold of 0.5 was set for the red light, signaling low-quality acquisitions that cannot be submitted and must be revised. Scores between 0.5 and 0.8 receive a yellow light, which allows submission but encourages further refinement if possible. These thresholds reflect our chosen criteria but can be adjusted as needed, which might lead to variations in quality outcomes and task completion rates and times.

4.2 Worker Recruitment

Worker recruitment was conducted via MicroWorkers.com, with two separate campaigns that had the same number of workers ($n_w = 50$) for both the control and test group. Both groups received the same task instructions: five image sections were shown to each worker, with the instructions to create a complete polygon around a tree. To maintain consistency, a sample image was provided to guide workers on acquisition standards.

All 50 image sections of Dataset B were processed, with each section processed by five different workers, resulting in 5 acquisitions per section ($n_a = 5$), where the same tree was processed, and thereby in a total of $50 \cdot n_a = 250$ acquisitions per group. As mentioned, five acquisitions were bundled into a single crowdworker job ($n_j = 5$), resulting in a total of 50 jobs per group, with each job performed by a single crowdworker. Compensation was standardized at $0.15 per job, ensuring equal incentives across both groups. Given 250 acquisitions per group, the total cost per group amounted to $37.50, resulting in an overall cost of $75 for the data acquisition.

4.3 Acquisition Examples

Following data collection through the described campaigns, the results were visualized to assess compatibility with Dataset B. Figure 5 provides an example image section of Dataset B, showing polygon acquisitions from both the test group, which had CNN-based feedback, and the control group, which did not.

Figure 5. Image section with 5 acquisitions: control group (left) and test group (right); green lines show ground truth.

Judging from Figure 5, the test group acquisitions (left) deliver more precise and consistent outlines around the center trees. In contrast, the control group acquisitions (right) show greater variability in polygon shapes and overall accuracy, potentially reflecting the absence of real-time feedback. While these initial observations suggest a quality difference, they do not quantify it, necessitating a comprehensive quality analysis with and without an integration process to confirm these observations.

5. Initial Quality Evaluation

We compared the IoU scores in respect to the reference between the control group, which didn’t have CNN feedback, and the test group, which used it. These results directly indicate a potential improvement of quality and consistency in the polygon acquisitions, as can be seen in Table 1.

	Control Group	Test Group
Mean	0.73	0.79
Standard Deviation	0.17	0.09

Table 1. Mean and standard deviation of IoU for control group and test group.

The control group, which worked without any guidance from the CNN, achieved a mean IoU of 0.73 and a standard deviation of 0.17. While the mean value indicates reasonably good results, the high standard deviation suggests that a notable portion of workers produced lower-quality acquisitions, a common occurrence in paid crowdsourcing, as discussed in the introduction. In contrast, the test group achieved a higher mean IoU of 0.79, combined with a much lower standard deviation of 0.09. This not only shows an improvement in accuracy, but also a notable improvement in terms of worker consistency.

The violin plot in Figure 6 highlights the differences in IoU distributions between the two groups. The width represents the frequency of IoU scores at each level in form of a kernel density estimate (KDE), while the center line is a standard boxplot.

As can be seen from Figure 6, the control group’s distribution is rather spread out, with a visible KDE in the lower IoU range. In contrast, the test group’s IoU scores are centered around higher values, with less deviations overall, inherently confirming the observed lower standard deviation. The histograms of both groups, that are visualized in Figure 7 along with their respective KDEs, further illustrate the observed differences.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-197-2025 | © Author(s) 2025. CC BY 4.0 License.

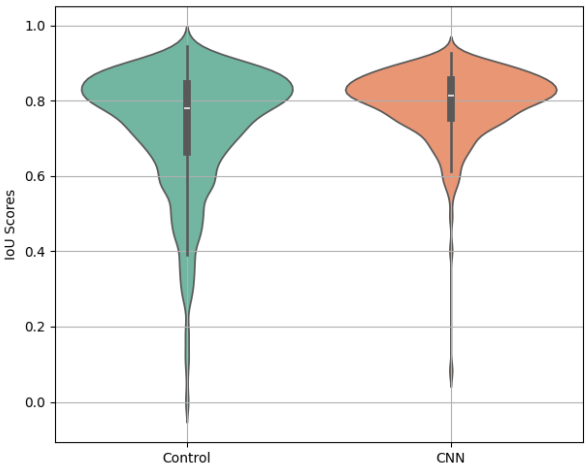

Figure 6. Violin plot of IoU scores for control group and test group.

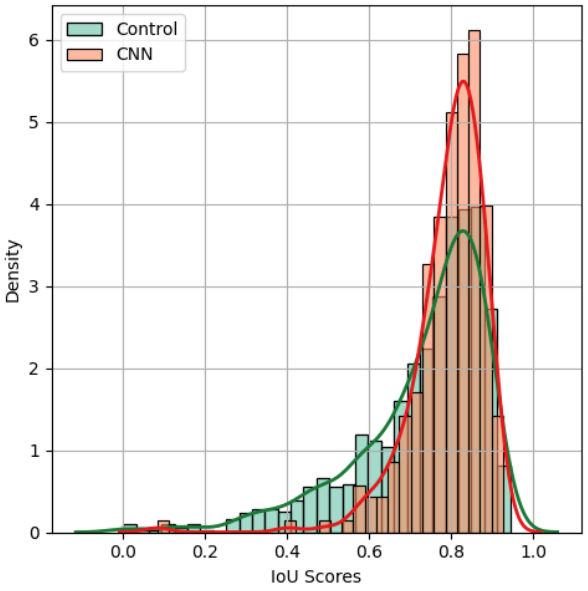

Figure 7. Histogram of IoU scores and corresponding KDE for control group and test group.

These visualized results demonstrate the impact of the real-time CNN feedback on the quality and consistency of polygon acquisition: For the control group, IoU scores are widely spread, with a notable portion of scores around or below 0.6, indicating cases of low-quality acquisitions. The CNN group, however, shows most scores clustered in the 0.7 to 0.9 range, with few scores falling below 0.6. For the higher end, i.e., IoU values above 0.9, the test group performed better than the control group as well. These results confirm the effectiveness of real-time feedback for crowdworkers across all IoU ranges, which is consistent with the objectives outlined in the methodology and with previous research, such as (Dow et al., 2012), is consistent.

Table 2 summarizes the percentage of acquisitions within each 0.2 IoU range for both groups and provides a clearer overview of this shift in IoU.

As shown in Table 2, we divided the possible IoU from 0 to 1 into intervals of steps with the size 0.2 to analyze quality distribution across different accuracy ranges. In the control group, a significant portion of acquisitions fell into the 0.4 - 0.6 range, which roughly corresponds to a red light in our system. Several acquisitions even scored below 0.4, underlining the lack of guidance and the resulting variability in quality. In contrast, the CNN-assisted group had close to none acquisitions below 0.4 and only a very small portion of 2.86% between 0.4 and 0.6, clearly showing improvement when compared to the control group. The test group scores slightly better results for the 0.6 - 0.8 range. Most notably, the test group showed a notable increase in high-quality acquisitions within the green light threshold (above 0.8), with over 57% of acquisitions achieving this level compared to only 45% in the control group. Subsequently, the test group performed strictly better for all IoU ranges. Furthermore, the more centered distribution improves both predictability and consistency, again underlining the benefit of our real-time feedback approach: Without external feedback, as observed in the control group, there is a notable risk of lower accuracy and a broader range of outputs.

IoU Range	Control Group (%)	Test Group (%)
0.0 - 0.2	1.24	0.41
0.2 - 0.4	2.07	0.00
0.4 - 0.6	14.05	2.86
0.6 - 0.8	37.19	39.59
0.8 - 1.0	45.45	57.14

Table 2. Percentage of acquisitions within IoU intervals of 0.2.

In summary, these findings demonstrate that the CNN-based approach with real-time feedback can notably improve both the accuracy and consistency of crowdsourced data. This higher output quality, along with increased predictability, potentially enables a reduction in redundant data acquisition, saving both time and money, and thereby improving scalability.

6. Integration & Redundancy

It has been established that the CNN-based approach can achieve higher output quality than traditional methods, suggesting that a reduction in redundant data acquisition may be possiblee. Additionally, we aim to examine the influence of integration, as described in (Collmar et al., 2023). Such an integration not only enhances quality through inherent majority voting but also consolidates results into a single output shape.

In order to assess the potential after integration, the integration was performed for $n_a = 1, \ldots, 5$ for both the control group and the test group. Mean and standard deviation values were calculated for each group, as shown in Table 3. To further highlight the efficiency of the CNN-assisted approach, the control group was extended to $n_a = 25$ while the test group remained unchanged. This allows for a comparison between the test group and a control group with a higher level of redundancy.

As can be seen from Table 3, an upward trend is visible for the IoU values, along with a downward trend for standard deviation values across both groups. This effect becomes noticeable starting from $n_a = 3$, where random effects due to particularly high or low-performing workers are minimized. Figure 8 visually illustrates these trends by plotting the mean IoU scores and displaying the standard deviations vertically around the data points for both groups across various sample sizes, i.e., choices of n_a.

When looking at both Table 3 and Figure 8, the improved data quality observed in the previous sections is also clearly evident after integration. For instance, the highest mean IoU for the

n_a	Mean (Ctrl)	Mean (Test)	Std (Ctrl)	Std (Test)
1	0.750	0.815	0.183	0.063
2	0.748	0.813	0.172	0.066
3	0.763	0.818	0.156	0.062
4	0.774	0.824	0.147	0.059
5	0.784	**0.827**	0.137	0.056
6	0.793	N/A	0.130	N/A
7	0.800	N/A	0.124	N/A
8	0.806	N/A	0.119	N/A
9	0.811	N/A	0.115	N/A
10	0.814	N/A	0.111	N/A
...	...	...	...	...
15	**0.827**	N/A	0.990	N/A
20	0.835	N/A	0.910	N/A
25	0.841	N/A	0.085	N/A

Table 3. Mean and standard deviation for different n_a.

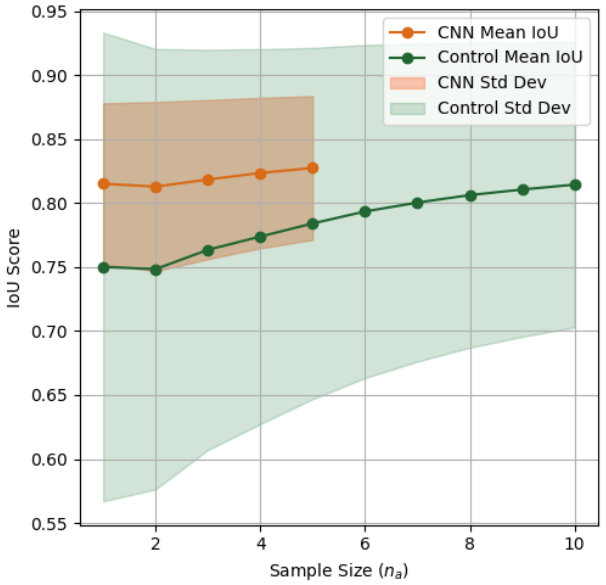

Figure 8. Mean and standard deviations for different n_a, shown only up to 10 for better readability.

CNN-assisted group is recorded at $n_a = 5$, reaching approximately 0.827, while the control group with the same sample size has a noticeably lower IoU of 0.784. Even when the control group's sample size is doubled, its mean IoU remains lower, only approaching the performance of the CNN-assisted group when n_a is increased threefold to 15. However, the standard deviation for $n_a = 15$ in the control group is nearly twice as high as the standard deviation for $n_a = 5$ in the CNN-assisted group. Even when n_a is increased to 25 for the control group, the standard deviation of the test group is still notably better compared to the control group, regardless of the chosen sample size for the test group. This highlights that the CNN group achieves both higher accuracy and consistency with less redundancy also after integration. Furthermore, even with just a single acquisition per image section, i.e., $n_a = 1$, the mean IoU for the CNN-assisted test group was comparable to the mean IoU results for $n_a = 10$ in the control group, resulting in a tenfold reduction in the number of necessary acquisitions. However, a single acquisition per image section may be too limited and prone to error due to small variations, which is why we recommend choosing $n_a \geq 3$.

In terms of cost reduction, both groups, i.e., the control group and the test group, are paid the same salary of \$0.15. The financial costs for data acquisition at $n_a = 5$ amount to \$37.50 each, as mentioned before. However, to achieve a similar output data quality in terms of IoU values, n_a had to be set to 15 for the control group, resulting in costs of $n_a = \$112.50$. In contrast, the CNN-assisted group, which incurred costs of only \$37.50, delivered significantly better results in terms of standard deviation, a level of performance that the control group could not achieve even for a much larger sample size.

It remains to add that, although there are costs in terms of money and time associated with training the CNN, these are one-time expenses. Particularly when considering the unmatched standard deviation observed in the test group, even at low sample sizes compared to high sample sizes in the control group, these one-time costs become secondary. Not only can overall costs be reduced to achieve the same mean level of IoU, but the results obtained through real-time CNN feedback also contribute to improved scalability for large datasets by providing acquisitions of better or similar quality with fewer annotations.

To illustrate this, Figure 9 shows a comparison of cumulative costs between a CNN-enhanced acquisition process, which includes the one-time training cost, and a traditional approach without real-time feedback, assuming both methods achieve the same quality level. The numbers used are those described in Section 4, which details the experimental setup: each job is paid \$0.15 for 5 polygons in both the test and control groups. Additionally, training data were acquired at \$0.10 for 5 polygons (or \$0.02 per acquisition), resulting in total training costs of approximately \$300. The figure visualizes different potential worker ratios for n_a between the control and test groups, derived from the results in Table 3. The actual worker ratio may vary by application, so we present multiple lines as general solutions.

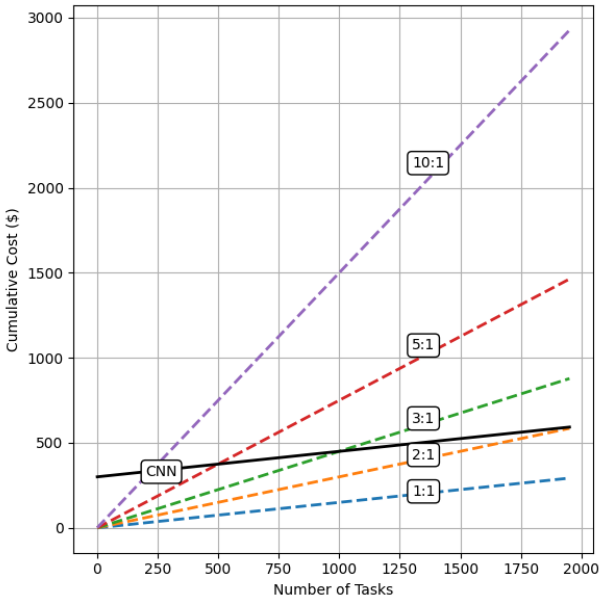

Figure 9. Cumulative cost comparison for different worker ratios.

While the CNN approach initially incurs higher costs due to the necessary training, the cumulative cost grows more slowly over time as the reduced sample size leads to lower costs per acquisition. As the number of tasks increases, the CNN-enhanced approach yields significant savings for the same data quality.

This makes it a more scalable and cost-effective solution for large-scale data collection campaigns: For the case of a worker ratio of 3:1, as in this study ($n_a = 15$ vs. $n_a = 5$ in Table 3), a sample size of 1,000 tasks or more is sufficient to achieve cost savings. In contrast, with a worker ratio of 10:1 ($n_a = 10$ vs. $n_a = 1$ in Table 3), cost savings appear much sooner, presenting even greater opportunities for efficiency and scalability.

7. Limitations & Future Research

While we have demonstrated and discussed the strengths of our approach, several limitations and areas for improvement remain to be considered.

To begin with, the datasets themselves: Since this study serves as a proof-of-concept, the focus on tree outlines rather than other objects appears justifiable. The training of our CNN relied on a single dataset with limited diversity and environmental variations such as lighting, weather or seasonal changes. While the model delivered strong results on the second dataset, incorporating additional datasets with diverse environmental conditions during training could further enhance generalization and build upon the already solid results.

The CNN itself was designed to be lightweight in order to both maximize the processing speed to handle acquisitions from multiple crowdworkers in parallel, while still maintaining low latency and accurate results. An alternative approach, although potentially influencing scalability, could reduce the number of parallel workers in order to allow more resources to be allocated to each instance of the CNN. Therefore, more resources would be available to focus on optimizing feature extraction and overall accuracy of the CNN, which could involve adding additional layers after the concatenation of the two branches, using the previously discussed multi-layer CNN, or even incorporating transformer architectures. Ultimately, the primary goal of this study was to demonstrate the benefit of real-time feedback, with the specific CNN structure being of secondary importance.

Separately, the CNN was trained using IoU as the primary quality metric. While IoU is a standard choice, it doesn't necessarily capture fine geometric details but rather an overall impression of the delivered polygon. A combination of parameters, such as the higher polygon moments, as used for the filtering process in (Collmar et al., 2023), could be used instead, allowing for a more detailed real-time analysis and potentially further increasing output quality. Similarly, in the analysis of the crowdworkers' resulting polygons, additional metrics such as Hausdorff distance could capture these fine geometric details better and subsequently deliver a more comprehensive view of the submitted polygons' quality.

Furthermore, the redundancy analysis could be further extended. Our results show that increasing the number of acquisitions per image section led to higher mean IoU values and reduced standard deviations, up to our tested maximum of five acquisitions. However, this raises the question whether these trends would persist with larger sample sizes, and if so, what sample size might lead to a potential saturation point or provide the most cost-efficient solution.

Future work could include further refining the proposed approach while keeping costs, both in terms of time and money, low. Of the previously mentioned points, we consider replacing or supplementing the IoU quality metric for the CNN training to be the most important. Instead of rating the overall polygonal shapes, as is done via the used intersection of union, local features and details might be essential as well. Consequently, including alternative parameters that capture local variations or provide a more detailed description of polygons could lead to a further enhanced model accuracy.

Additionally, future work could involve expanding the current traffic-light feedback system by a more sophisticated approach. While the current feedback system provides three levels (poor, average, or high quality), adding more options or allowing crowdworkers to respond and provide feedback could help create a more adaptive system. Furthermore, tracking how real-time feedback affects the crowdworkers' engagement and task efficiency over time could also help to optimize feedback mechanism. It is unclear, however, whether a more sophisticated feedback system, such as error bars or potentially even gamification elements like high scores, would further improve data quality or overly complicate the process, as crowdworkers may have only a few seconds to interpret the feedback. In any case, further research might allow to optimize our proposed approach even further.

8. Conclusion

We proposed a real-time feedback mechanism to improve data quality in paid crowdsourcing, with a proof-of-concept approach for the acquisition of geometric tree outlies. We used a lightweight two-branch CNN for a scalable solution with minimal latency and high performance even for multiple clients in parallel. The use of different datasets indicated potential generalization across different environmental effects. The inclusion of our traffic light feedback system resulted in better data quality compared to the reference, as measured by intersection over union of resulting shapes. Not only was the average data quality higher, but data distribution also improved notably, enabling a reduction in sample sizes for redundant data acquisitions while maintaining, or even enhancing, output quality.

When an integration process is performed, the gap in data quality widens even further, achieving a standard deviation that would otherwise require notably larger sample sizes. This approach can be leveraged in two ways: to significantly increase data quality at the same cost for applications where quality is the primary concern, or to reduce sample sizes for comparable or improved results, allowing for cost savings in time and money for large-scale datasets, where the one-time cost of network training is feasible.

Acknowledgements

Partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2120/1 – 390831618.

References

Brabham, D. C., 2013. *Crowdsourcing*. MIT Press, Cambridge, MA, USA.

Budde, L. E., Collmar, D., Sörgel, U., Iwaszczuk, D., 2024. Investigating the relationship between image quality and crowdsourcing for labeling. 44. Wissenschaftlich-Technische Jahrestagung der DGPF.

Chandana, R., Ramachandra, A., 2022. Real time object detection system with YOLO and CNN models: A review. *arXiv Prepr. arXiv2208*, 773.

Chandler, J., Paolacci, G., Mueller, P., 2013. Risks and rewards of crowdsourcing marketplaces. *Handbook of human computation*, Springer, 377–392.

Chen, X., Li, D., Liu, M., Jia, J., 2023. CNN and Transformer Fusion for Remote Sensing Image Semantic Segmentation. *Remote Sensing*, 15(18), 4455.

Collmar, D., Walter, V., Kölle, M., Sörgel, U., 2023. From Multiple Polygons to Single Geometry: Optimization of Polygon Integration for Crowdsourced Data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 159–166.

Collmar, D., Walter, V., Soergel, U., 2024. Crowd Controls Crowd: Quality Improvement of Polygon Integration in Paid Crowdsourcing. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 83–90.

Cui, L., Chen, J., He, W., Li, H., Guo, W., Su, Z., 2021. Achieving approximate global optimization of truth inference for crowdsourcing microtasks. *Data Science and Engineering*, 6(3), 294–309.

Dow, S., Kulkarni, A., Klemmer, S., Hartmann, B., 2012. Shepherding the crowd yields better work. *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 1013–1022.

Hirth, M., Hoßfeld, T., Tran-Gia, P., 2011. Anatomy of a crowdsourcing platform-using the example of microworkers. com. *2011 Fifth international conference on innovative mobile and internet services in ubiquitous computing*, IEEE, 322–329.

Hossain, M., 2012. Users’ motivation to participate in online crowdsourcing platforms. *2012 International Conference on Innovation Management and Technology Research*, IEEE, 310–315.

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.

Kobayashi, M., Morita, H., Morishima, A., 2022. Efficient crowdsourcing for semantic segmentation considering human cognitive characteristics. *International Conference on Human-Computer Interaction*, Springer, 300–307.

Liu, X., Ghazali, K. H., Han, F., Mohamed, I. I., 2023. Review of CNN in aerial image processing. *The Imaging Science Journal*, 71(1), 1–13.

Microworkers, 2024. Microworkers: A crowdsourcing platform. https://www.microworkers.com. Accessed: 2024-06-19.

Ostyakova, L., Smilga, V., Petukhova, K., Molchanova, M., Kornev, D., 2023. Chatgpt vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 242–254.

Ozdemir, M. A., Elagoz, B., Alaybeyoglu, A., Sadighzadeh, R., Akan, A., 2019. Real time emotion recognition from facial expressions using cnn architecture. *2019 medical technologies congress (tiptekno)*, IEEE, 1–4.

Rasmussen, C. B., Kirk, K., Moeslund, T. B., 2022. The challenge of data annotation in deep learning—a case study on whole plant corn silage. *Sensors*, 22(4), 1596.

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., Vukovic, M., 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. *Proceedings of the international AAAI conference on web and social media*, 5(1), 321–328.

Shi, W., Meng, F., Wu, Q., 2017. Segmentation quality evaluation based on multi-scale convolutional neural networks. *2017 IEEE Visual Communications and Image Processing (VCIP)*, IEEE, 1–4.

Steier, J., Goebel, M., Iwaszczuk, D., 2024. Is your training data really ground truth? A quality assessment of manual annotation for individual tree crown delineation. *Remote Sensing*, 16(15), 2786.

Vuurens, J., de Vries, A. P., Eickhoff, C., 2011. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR’11)*, 21–26.

Wang, X., Chen, L., Ban, T., Lyu, D., Guan, Y., Wu, X., Zhou, X., Chen, H., 2023. Accurate label refinement from multiannotator of remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–13.

Zhang, J., 2022. Knowledge learning with crowdsourcing: A brief review and systematic perspective. *IEEE/CAA Journal of Automatica Sinica*, 9(5), 749–762.

Zhang, J., Wu, X., Sheng, V. S., 2016. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*, 46, 543–576.

Zheng, Y., Li, G., Li, Y., Shan, C., Cheng, R., 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5), 541–552.

Zhu, D., Carterette, B., 2010. An analysis of assessor behavior in crowdsourced preference judgments. *SIGIR 2010 workshop on crowdsourcing for search evaluation*, 17–20.