Weak Segmentation and Unsupervised Evaluation: Application to Froth Flotation Images

Egor Prokopov, Daria Usacheva, Mariia Rumiantceva[a] and Valeria Efimova[b]

ITMO University, Kronverkski prospect,49, Saint-Petersburg, Russia

egorprokopov216@gmail.com, obobojk@gmail.com, {marrum, vefimova}@itmo.ru

Keywords: Froth Flotation, Image Segmentation, Foundation Models, Weakly-Supervised Learning, Unsupervised Evaluation.

Abstract: Images featuring clumped texture object types are prevalent across various domains, and accurate analysis of this data is crucial for numerous industrial applications, including ore flotation—a vital process for material enrichment. Although computer vision facilitates the automation of such analyses, obtaining annotated data remains a challenge due to the labor-intensive and time-consuming nature of manual labeling. In this paper, we propose a universal weak segmentation method adaptable to different clumped texture composite images. We validate our approach using froth flotation images as a case study, integrating classical watershed techniques with foundational models for weak labeling. Additionally, we explore unsupervised evaluation metrics that account for highly imbalanced class distributions. Our dataset was tested across several architectures, with Swin-UNETR demonstrating the highest performance, achieving 89% accuracy and surpassing the same model tested on other datasets. This approach highlights the potential for effective segmentation with minimal manual annotations while ensuring generalizability to other domains.

1 INTRODUCTION

Clumped texture composite images present unique challenges in image segmentation. Clumped textured data often refers to datasets where values or patterns tend to cluster together, making it challenging to analyze or interpret due to overlapping features, high density in specific regions, or non-uniform distribution. For example, in the industrial field, clumped textured data can include ore rocks on a conveyor belt, pores in gypsum boards, pellets in metallurgical processes, animals clustered in enclosures, and cells grouped in microscopic images for quality control or analysis. Froth flotation bubbles represent an excellent example of clumped textured data, as they exhibit dense, irregular clustering patterns that are critical for analyzing and optimizing separation processes in mining and mineral processing industries.

Froth flotation is an ore enrichment process that separates hydrophobic particles from hydrophilic ones at the interface between phases. The quality of froth is critical for effective control of the flotation process, with visual characteristics such as bubble quantity, shape, texture, and color playing significant roles in determining quality (Gui et al., 2013; Saghatoleslam et al., 2004; Subrahmanyam and Forssberg, 1988). While technologists often rely on visual evaluations, these methods can be subjective and inaccurate due to the dynamic nature of froth.

Computer vision techniques can automate the estimation of flotation parameters, but they typically require large labeled datasets for accurate segmentation—a task that becomes particularly challenging in the context of clumped froth bubbles, as well as other similar classes of data. Simple detection methods that identify centroids fail to provide sufficient information regarding object boundaries, shapes, or sizes, limiting their effectiveness across various applications. Furthermore, manually labeling images is time-consuming and prohibitively expensive, given the complexities of froth structures and the diverse nature of other object classes.

To address the high costs of manual labeling for clumped texture data, we propose a weak segmentation approach that generates pseudo-labels for froth flotation images. Using this method as a case study, we trained several segmentation networks on a created dataset, with the aim of estimating froth features and informing other algorithms, such as those control-

[a] https://orcid.org/0000-0002-0988-2762
[b] https://orcid.org/0000-0002-5309-2207

ling flotation machinery.

To evaluate the stability and robustness of our weak labeling method, we employed an unsupervised temporal consistency metric (Varghese et al., 2020) and introduced a new optical flow similarity metric. Additionally, we utilized the object recall metric to quantify segmentation completeness. Our approach offers a viable alternative for analyzing various data types beyond froth flotation without the need for extensive labeled datasets.

2 RELATED WORKS

2.1 Clumped Data and Froth Bubbles

In many industrial applications, data contain numerous small objects that pose significant challenges for image segmentation due to clustering and homogeneity in textures, shapes, or colors. This difficulty is compounded by the low quality of images, which are often degraded by noise, blurring, or inconsistent lighting.

For example, datasets in materials processing industries may include aggregates, pellets, or stones (fengkai, 2024; Pellet, 2023; Anton, 2024), where minimal visual variance and object overlap make segmentation difficult. Flotation bubbles, another form of clumped texture data, present similar challenges. Widely studied in industries like mineral processing, flotation bubbles share similar shapes and textures, complicating their individual distinction (Moolman et al., 1995). Manual annotation of these datasets is difficult, leading to the exploration of various segmentation methods such as the watershed algorithm (Peng et al., 2021), U-Net (Ju et al., 2022; Zhang and Xu, 2020), and classification models (Cao et al., 2021). The watershed algorithm is effective for touching objects with clear edges, while U-Net captures both local and global features for complex industrial images. These challenges highlight the need for innovative segmentation approaches, especially where precise annotations are expensive or time-consuming (Rumiantceva and Filchenkov, 2022).

2.2 Unsupervised and Weakly-Supervised Segmentation

Unsupervised segmentation methods aim to overcome the challenges of manual annotation in clumped texture data by identifying patterns without labeled data. Traditional techniques like the watershed algorithm (Peng et al., 2021) are effective but often struggle with high-density object regions and poor image quality.

Supervised segmentation, particularly with neural network architectures like UNet (Ronneberger et al., 2015) and Mask R-CNN (He et al., 2018), has been successful in handling clumped texture data. UNet's ability to capture fine details makes it popular in industrial applications (Zhong et al., 2023), though these models require large annotated datasets. The Segment Anything Model (SAM) (Kirillov et al., 2023) offers a significant advancement by leveraging a promptable segmentation framework, excelling in zero-shot transfer to new tasks, making it ideal for industries with frequent segmentation needs.
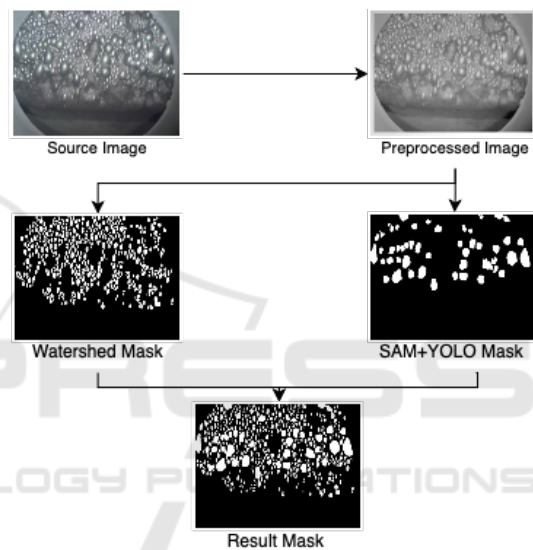

Figure 1: Scheme of the proposed weak segmentation method.

2.3 Metrics

Evaluating weakly-supervised methods is challenging due to the lack of labeled datasets, as conventional metrics like Intersection over Union (IoU) or pixel-wise accuracy require ground truth data. To overcome this, alternative metrics must be used. One such metric is temporal consistency (Varghese et al., 2020), which assesses segmentation stability across consecutive frames or images. It evaluates how consistent the segmentation results are when objects undergo slight variations in position, scale, or orientation, making it useful for video data or time-sequenced images.

Temporal consistency provides an indirect measure of a model's reliability, rewarding methods that maintain stable segmentations. Other metrics, such as object coherence and size regularity, can also offer approximate evaluations by ensuring segmented objects

maintain logical shapes and proportions. These approaches enable performance assessment in weakly-supervised settings without the need for ground truth.

3 METHOD

In this section, we describe our method for weak labeling of froth images, combining foundation models with classical computer vision techniques. Previous works have shown that segmenting clumped, clumped texture objects like froth bubbles is challenging, particularly due to poor image quality and the need for detailed annotations. Traditional methods, such as watershed and U-Net, require extensive labeled datasets, which are difficult to obtain. To address this, we propose a hybrid approach. We use SAM to extract big and medium size bubbles with guidance from YOLOv8 (Jocher et al., 2023). To segment small bubbles, we use watershed algorithm with morphological transformations. Finally, we apply post processing steps (Fig. 1). This allows us to minimize manual annotation while achieving effective segmentation through weak labeling and unsupervised learning.

3.1 Watershed Approach

First, the input image undergoes preprocessing to correct uneven lighting, using the Single-Scale Retinex (SSR) (Land and McCann, 1971) algorithm for normalization (Fig. 2.. Subsequently, bilateral filtering smooths the image while preserving the histogram's structure. Without SSR, the histogram would be skewed toward darker values, complicating further processing. Morphological transformations are then applied to enhance contrast between highlights and bubble regions, adjusting the histogram to facilitate optimal threshold selection for segmentation.
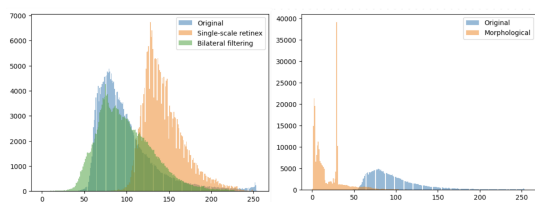

Figure 2: Histogram of the image after applying steps of the preprocessing.

As a result, we have successfully extracted the bubble markers. These markers enable the assessment of relative size, velocity, and quantity of the bubbles. For analytical purposes, we categorize the bubbles into two groups: small bubbles, and medium to large bubbles.
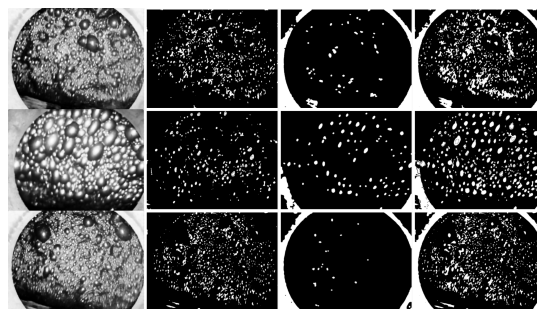

Figure 3: The first column is the original image, the second one corresponds to extracted markers of small sized bubbles, the third one is big and medium sized bubbles markers, the last one is the watershed output.

Based on these highlighted regions, we generate the seed areas for the watershed algorithm. A combination of erosion and dilation, with varying kernel sizes, is applied to preprocess the image. The resulting processed image is then used as input for the watershed algorithm, producing contours and segmentation masks. This approach proves effective in generating masks for small-sized bubbles, although the results may be less precise. However, due to the substantial variability in the shapes of medium and large bubbles, this method is not suitable for accurately segmenting these categories(Fig. 3, Fig. 4).
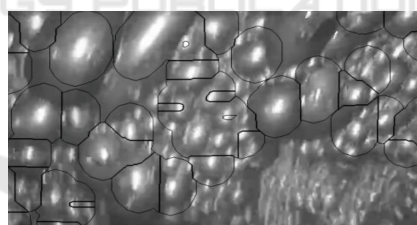

Figure 4: The proposed method of watershed processing does not segment big bubbles.

3.2 Foundation Model Approach

To segment medium and large bubbles in flotation froth, we used the Segment Anything Model (SAM) (Kirillov et al., 2023). As a foundation model, SAM does not require additional training to label the data. However, using SAM without guidance does not yield satisfactory results.
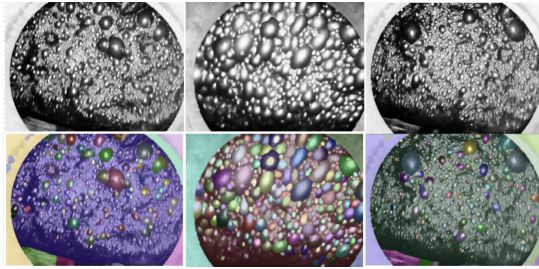
Figure 5: Output of SAM without any prompt guidance. The first row is the input image, the second is the outputs of the model.

As shown in Fig. 5, SAM misses most of the bubbles and incorrectly segments objects unrelated to froth flotation. To achieve the desired results, guidance is necessary. SAM supports sparse prompts (point, bounding box, and text) as well as dense prompts (mask). Since our dataset lacks labeled masks, dense prompts could not be used. Furthermore, text prompts failed to address the issues observed with unguided SAM. Thus, we considered two guidance approaches: bounding box prompts and a combination of bounding boxes and point prompts.

To detect bubbles in flotation images and generate bounding boxes, we employed the YOLOv8 detection model (Fig. 6). This model was trained on a small hand-labeled dataset of 70 images. Notably, precise labeling was unnecessary for this task, as some small bubbles are difficult to recognize even for experts, especially when image quality is poor. These bounding boxes served as guidance for SAM to enhance segmentation accuracy.
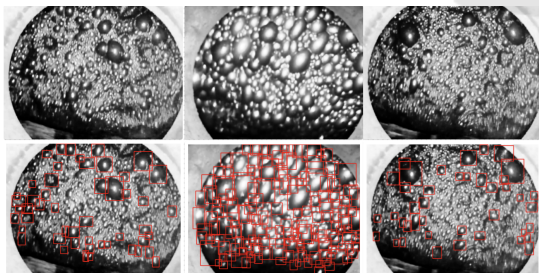

Figure 6: Output of YOLOv8 detection model for medium and large bubbles. The first row is the original image, the second row is the visualization of the predicted bounding boxes.

We evaluated two approaches for prompting SAM. The first approach used only bounding boxes as prompts for segmentation, while the second approach added point prompts derived from the centers of each bounding box. Among these, using only bounding box prompts produced cleaner and less noisy masks, as shown in Fig. 7.
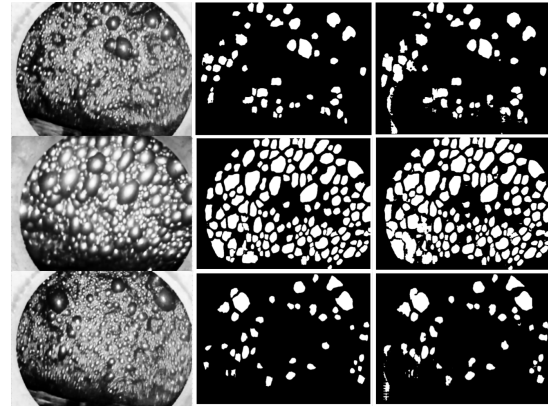

Figure 7: Output of SAM without point prompt and with it. The first column is raw data, the second one is output of the model using only bounding boxes, the third one is with additional point guidance.

3.3 Post-Processing

At this step of the proposed method, we post-process the results obtained from the previous steps to generate a single comprehensive mask. Initially, we utilize a set of masks for each bubble derived from the Segment Anything model, which are then narrowed by a specific factor—determined to be 0.2 through experimentation. These individual masks are subsequently combined into a unified mask. The output from the watershed processing is then incorporated, resulting in the final segmentation label (Fig. 8).

4 EXPERIMENTS

In this section, we present the results of training segmentation neural networks on the dataset labeled using our method. We compare the robustness and stability of the masks generated by the watershed algorithm, our approach, and the discussed neural networks. Additionally, we introduce a novel metric to evaluate the stability of segmentation across contiguous frames, which is based on the similarity between the optical flows of the frames and their corresponding masks.

4.1 Dataset Development

To develop the dataset, we utilized footage from various flotation machines, selecting every tenth frame from the provided videos. During preprocessing, we manually removed duplicate frames and excluded low-quality images that contained codec errors and artifacts. Each video was labeled independently, and the resulting labeled frames were subsequently merged

into training, validation, and test sets without shuffling frames from different videos. Using this method, we generated a total of approximately 15,000 labeled images.
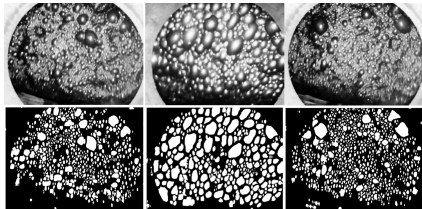

Figure 8: Result of proposed method. The first row is the input image, the second row is the result of the proposed algorithm.

4.2 Model Training

We utilized our dataset to train several segmentation models, including HRNet, DeepLabv3, Segformer, and Swin with UNETR decoder. The rationale behind the selection of these models was to evaluate the learning outcomes of various contemporary architectures:

- The **HRNet** (Wang et al., 2020) architecture utilizes features at various semantic levels, known for delivering informative and accurate segmentation results.

- The **DeepLabv3** (Chen et al., 2018) model employs sparse convolutions with varying kernel sizes to extract features across different semantic levels, making it lightweight and suitable for real-time processing.

- The **Segformer** (Xie et al., 2021) is a vision transformer model that enhances generalization through its attention mechanism, requiring substantial data for effective training.

- **Swin-UNETR** (Hatamizadeh et al., 2022) adapts the classic UNet architecture with residual connections in the decoder, using Swin Transformer (Liu et al., 2021) as the encoder.

Table 1: Supervised metrics table.

Model	IoU	Dice	Acc.	Prec.	Rec.
HRNet	0.40	0.58	0.88	0.64	**0.52**
DeepLabv3	0.38	0.55	0.85	0.70	0.45
SegFormer	0.38	0.56	0.86	0.68	0.49
Swin-UNETR	**0.48**	**0.63**	**0.89**	**0.86**	0.51

Prior to training, we preprocessed the input images utilizing the CLAHE algorithm to ensure consistent lighting (Mishra, 2021). Additionally, we applied a variety of data augmentation techniques, including random flipping, cropping, translation, scal-

ing, and rotation. To mitigate data leakage, the dataset was initially partitioned into training, validation, and test subsets according to the source videos. Consequently, we assessed the performance of the trained models on the test dataset. As presented in Table 1, Swin-UNETR exhibited the best overall performance, achieving the highest Intersection over Union (IoU), Dice coefficient, accuracy, and precision. Although HRNet demonstrated a marginally better recall, the balanced performance of Swin-UNETR positions it as the most effective model for our segmentation task. It is noteworthy that the model should avoid overfitting to the weak labels, which accounts for the relatively modest IoU and Dice values observed.

4.3 Metrics

To evaluate stability and robustness of froth images segmentation, we used temporal consistency metric. We also introduce optical flow similarity metric to avoid image warping step used in temporal consistency.

Considering that neither temporal consistency nor optical flow similarity metrics directly represent the number of objects estimated, we decided to include object recall as one of our metrics. Although recall is a supervised metric, we found a way to make it independent of labeled data by using watershed markers to count objects.

We compared the resulting masks from the watershed algorithm and the proposed method.

4.3.1 Temporal Consistency

Most supervised metrics assess segmentation accuracy but not its stability between contiguous frames. Additionally, such metrics require ground truth data, making them unsuitable for comparing weakly labeled segmentation algorithms. In (Varghese et al., 2020), an unsupervised method was introduced to estimate temporal consistency between contiguous frames using optical flow. The original method relied on Farnebäck's algorithm (Farnebäck, 2003), but classical optical flow methods perform poorly on clumped texture data. Instead, we used the RAFT model (Teed and Deng, 2020), a robust neural network-based approach implemented in PyTorch (Paszke et al., 2019).

Our analysis revealed that the temporal consistency metric is not well-suited for evaluating segmentation stability on clumped texture data. This is due to significant differences on image borders between the current frame mask and the transformed previous mask (Fig. 9), strongly affecting the IoU metric. These differences arise from froth motion and the appearance/disappearance of bubbles between

frames. As the original metric does not account for such changes, we modified it for our data by omitting the mask transformation step.
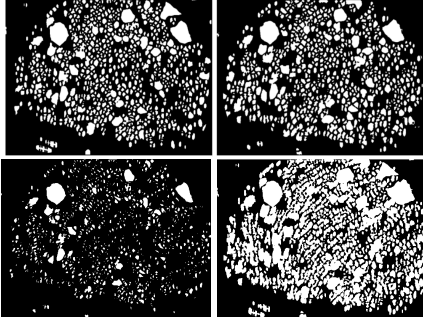

Figure 9: Comparison of the current frame mask (top-left), the transformed mask of the previous frame (top-right), intersection of the masks (bottom-left) and the union of the masks (bottom-right).

4.3.2 Optical Flow Similarity

To address the issue of instability in segmentation evaluation, we propose calculating the mean cosine similarity between the optical flow vectors of contiguous frames and their corresponding masks (Fig. 11), instead of using the intersection over union metric between the current frame mask and the transformed previous frame mask.

The rationale is simple: if segmentation is stable and robust, the mask's optical flow vectors should align with the frame's optical flow vectors. This approach minimizes the influence of optical flow inaccuracies on the final metric.

4.3.3 Object Recall

In comparing the results of our proposed method with the watershed segmentation approach (Fig. 10), we noted a marked discrepancy in the number of detected objects between the two techniques. To quantitatively assess this difference, we opted to utilize the recall metric. However, given that recall is inherently a supervised metric, we adapted it for unsupervised evaluation by anchoring our metric to the number of objects identified by an alternative algorithm.
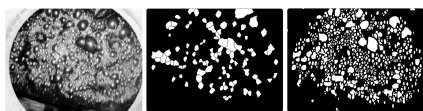

Figure 10: Comparison of the watershed and the modified watershed.

Initially, YOLOv8 was considered to be used for object counting. As labelling dataset for object de-

tection is far less labour intensive, we created dataset consisting of 20 images for fine-tuning the model. However, the results for estimating the number of objects were unsuitable as it missed a significant number of small bubbles (Fig. 12a). During training, it was observed that the centers of bounding boxes were located at the bubble glare spots. Given that the markers obtained from image preprocessing for the watershed algorithm effectively highlight the glare spots of all bubbles, particularly the small ones, we decided to use these markers for our metric.
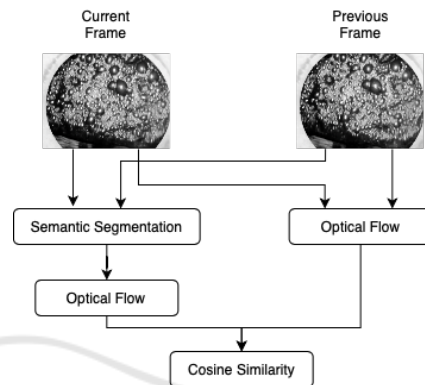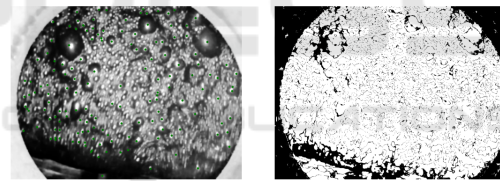

Figure 11: Scheme of the optical flow similarity metric.

(a) Found objects   (b) Dynamic mask

Figure 12: The resulting bounding boxes centers are marked with green dots and the dynamic mask of changes.

The centers of each marker are calculated with a focus on those involved in the flotation process, which is inherently dynamic. To compute the "static" component of the image, two contiguous frames are converted to grayscale, and the absolute difference between them is computed. Subsequently, static regions are filtered out using a very low threshold of 3 out of 255. This process results in an array where a value of 0 represents the static portion and a value of 1 denotes the dynamic component, which constitutes the rescaled mask (Fig. 12b). To eliminate false markers located in regions not exhibiting the flotation process, the markers are logically multiplied by the obtained mask. The total number of remaining markers represents $TP + FN$. The centers of the markers are logically multiplied by the segmentation masks obtained after processing with the aforementioned methods. The resulting number of markers is considered

TP, and Recall is calculated as $\frac{TP}{TP+FN}$.

Table 2: Unsupervised metrics table.

Method	Temp. Cons.	Opt. Flow Sim.	Obj. Rec.
Watershed	**0.41**	0.23	0.52
Ours	0.30	**0.27**	**0.85**

5 DISCUSSION

The proposed weak segmentation method demonstrates strong performance as measured by the optical flow similarity metric; however, it exhibits some instability that may hinder its robustness. While it outperforms the previous watershed segmentation approach with manual mask corrections by identifying more images, its occasional instability reduces these advantages. Improving the method's consistency would enhance its effectiveness compared to traditional techniques.

The quality of the original data significantly influences labeling accuracy. During analysis, several issues were identified: poor resolution, artifacts that interfere with segmentation, and irrelevant objects complicating the process. Our results revealed key inaccuracies, including a tendency to over-segment certain regions and overlook smaller objects, such as bubbles.

Furthermore, challenges arise when using unsupervised metrics. The temporal consistency metric, which compares transformed masks via the Intersection over Union (IoU), may not effectively apply to clumped texture data with numerous objects. Although the optical flow similarity metric avoids mask transformation, it remains sensitive to original data quality. Additionally, while our object recall metric successfully detects most objects, it does not guarantee comprehensive detection, rendering it an approximation for evaluation purposes.

To address these issues, generating synthetic clumped texture data presents a potential solution, aiming to alleviate challenges posed by low-quality data and reduce the need for extensive annotation. We propose that a synthetic dataset could serve as reliable ground truth for training neural networks. Moreover, combining weakly labeled masks produced by our method with synthetic data from three-dimensional modeling techniques could enhance training by providing realistic and diverse examples, thereby improving the robustness and accuracy of segmentation models.

6 CONCLUSION

In this work, we present a weak segmentation approach for froth flotation images, tackling key challenges in mineral processing. Our method combines classical computer vision techniques with advanced neural network architectures, specifically the Segment Anything Model (SAM) and the YOLOv8 model for bubble detection. This hybrid approach enhances the accuracy and efficiency of object identification in highly clumped froth images, which are traditionally difficult to segment.

A major contribution of our work is the development of a more precise weak labeling method that easily adapts to various types of clumped texture data, along with the creation of a dataset containing 15,000 images. This approach reduces the labor-intensive manual labeling process by minimizing the need for extensive annotated datasets. Notably, we have successfully tested the method in another domain, specifically on stones.
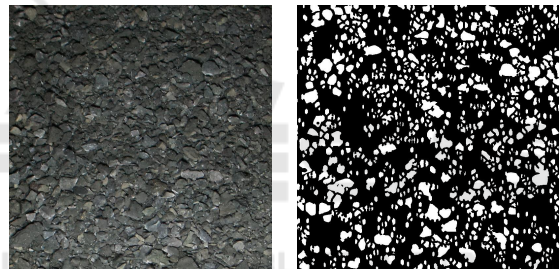

Figure 13: Image and mask for stones.

We also introduce unsupervised evaluation metrics for assessing the stability and robustness of segmentation through counting, utilizing temporal consistency and optical flow similarity metrics. While our method shows promise, we acknowledge its limitations related to instability and dependence on the original data quality. Future work will focus on generating synthetic clumped texture data to establish a more reliable ground truth for training neural networks, further enhancing the method's applicability in real-world scenarios. Notably, we have successfully validated this method on stone datasets, yielding positive results.

ACKNOWLEDGEMENTS

REFERENCES

Anton (2024). Rocks dataset. https://universe.roboflow.com/anton-yjhge/rocks-bhdzr. visited on 2024-10-22.

Cao, W., Wang, R., Fan, M., Fu, X., Wang, H., and Wang, Y. (2021). A new froth image classification method based on the mrmr-ssgmm hybrid model for recognition of reagent dosage condition in the coal flotation process. *Applied Intelligence*, 52(1):732–752. [Online; accessed 2023-10-15].

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion.

fengkai (2024). coal2.1 dataset. https://universe.roboflow.com/fengkai-ncemj/coal2.1. visited on 2024-10-22.

Gui, W., Liu, J., Yang, C., Chen, N., and Liao, X. (2013). Color co-occurrence matrix based froth image texture extraction for mineral flotation. *Minerals Engineering*, s 46–47:60–67.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., and Xu, D. (2022). Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2018). Mask r-cnn.

Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics YOLO.

Ju, Y., Wu, L., Li, M., Xiao, Q., and Wang, H. (2022). A novel hybrid model for flow image segmentation and bubble pattern extraction. *Measurement*, 192:110861. [Online; accessed 2023-10-15].

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. (2023). Segment anything.

Land, E. H. and McCann, J. J. (1971). Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1):1–11.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows.

Mishra, A. (2021). Contrast limited adaptive histogram equalization (clahe) approach for enhancement of the microstructures of friction stir welded joints.

Moolman, D. W., Aldrich, C., van Deventer, J., and Bradshaw, D. (1995). The interpretation of flotation froth surfaces by using digital image analysis and neural networks. *Chemical Engineering Science*, 50:3501–3513.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8024–8035.

Pellet (2023). Ensemble pellet dataset. https://universe.roboflow.com/pellet-xbnmm/ensemble-pellet. visited on 2024-10-22.

Peng, C., Liu, Y., Gui, W., Tang, Z., and Chen, Q. (2021). Bubble image segmentation based on a novel watershed algorithm with an optimized mark and edge constraint. *IEEE Transactions on Instrumentation and Measurement*, PP:1–1.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

Rumiantceva, M. and Filchenkov, A. (2022). Deep learning and pseudo-labeling for ore granulometry. *Procedia Computer Science*, 212:387–396. 11th International Young Scientist Conference on Computational Science.

Saghatoleslam, N., Karimi, H., Rahimi, R., and Shirazi, H. (2004). ... of texture and color froth characteristics for evaluation of flotation performance in sarcheshmeh copper pilot plant using image analysis and neural .... *IJE Transactions B*, 17.

Subrahmanyam, T. V. S. and Forssberg, E. (1988). Froth stability, particle entrainment and drainage in flotation : a review. *International Journal of Mineral Processing*, 23:33–53.

Teed, Z. and Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow.

Varghese, S., Bayzidi, Y., Bär, A., Kapoor, N., Lahiri, S., Schneider, J. D., Schmidt, N., Schlicht, P., Hüger, F., and Fingscheidt, T. (2020). Unsupervised temporal consistency metric for video segmentation in highly-automated driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1369–1378.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., and Xiao, B. (2020). Deep high-resolution representation learning for visual recognition.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers.

Zhang, L. and Xu, D. (2020). Flotation bubble size distribution detection based on semantic segmentation. *IFAC-PapersOnLine*, 53(2):11842–11847. [Online; accessed 2023-10-15].

Zhong, Y., Tang, Z., Zhang, H., Xie, Y., and Gao, X. (2023). A froth image segmentation method via generative adversarial networks with multi-scale self-attention mechanism. *Multimedia Tools and Applications*, 83:1–20.