Segmenting the complex and irregular in two-phase flows

A real-world empirical Study with SAM2

Küçük, Semanur; Della Santina, Cosimo; Laskari, Angeliki

DOI
10.1016/j.ijmultiphaseflow.2025.105557

Publication date
2026

Document Version
Final published version

Published in
International Journal of Multiphase Flow

Citation (APA)
Küçük, S., Della Santina, C., & Laskari, A. (2026). Segmenting the complex and irregular in two-phase flows: A real-world empirical Study with SAM2. *International Journal of Multiphase Flow*, *196*, Article 105557. https://doi.org/10.1016/j.ijmultiphaseflow.2025.105557

International Journal of Multiphase Flow 196 (2026) 105557

Contents lists available at ScienceDirect

International Journal of Multiphase Flow

journal homepage: www.elsevier.com/locate/ijmulflow

Express Track Article

Segmenting the complex and irregular in two-phase flows: A real-world empirical Study with SAM2

Semanur Küçük *, Cosimo Della Santina, Angeliki Laskari

Delft University of Technology, Faculty of Mechanical Engineering, Mekelweg 5, Delft, 2628 CD, Netherlands

ARTICLE INFO

Dataset link: https://www.kaggle.com/datasets/semanurkk/bubblyflow, https://github.com/Semanur97/BubblyFlow.git

Keywords:
SAM2.1
Bubble segmentation
Multiphase flows
Air lubrication
Transfer learning

ABSTRACT

Segmenting gas bubbles in multiphase flows is a critical yet unsolved challenge in numerous industrial settings, from metallurgical processing to maritime drag reduction. Traditional approaches — and most recent learning-based methods — assume near-spherical shapes, limiting their effectiveness in regimes where bubbles undergo deformation, coalescence, or breakup. This complexity is particularly evident in air lubrication systems, where coalesced bubbles form amorphous and topologically diverse patches. In this work, we revisit the problem through the lens of modern vision foundation models. We cast the task as a transfer learning problem and demonstrate, for the first time, that a fine-tuned Segment Anything Model (SAM v2.1) can accurately segment highly non-convex, irregular bubble structures using as few as 100 annotated images.

1. Introduction

Accurate segmentation of bubbles or air patches from optical measurements plays a crucial role in analyzing two-phase flows, as it underpins the study of drag reduction, turbulence modulation, and interfacial dynamics (Tanaka et al., 2023; Ni, 2024; Wang et al., 2023). However, this important task remains challenging due to several factors, including overlapping bubble boundaries, inconsistent lighting conditions, image noise, and irregular bubble shapes that deviate from ideal spherical forms. Qin et al. (2017) demonstrate that standard segmentation algorithms (Serra et al., 2020; Ronneberger et al., 2015; Schmidt et al., 2018; He et al., 2017) often fail to provide accurate results, forcing researchers to manually inspect and refine the segmentation, thus substantially limiting the scalability of the analyses.

To overcome these limitations and motivated by recent advances in computer vision, bubble detection research has increasingly shifted toward deep learning, in the hope that these techniques can better handle complex scenarios. Early efforts by Ilonen et al. and Serra et al. explored the application of flat ANN to bubble segmentation, establishing baselines for later data-driven methods (Ilonen et al., 2018; Serra et al., 2020). This was followed by the adoption of vanilla convolutional neural networks (CNNs), with Soibam et al. and Malakhov et al. targeting boiling flows under constrained conditions (Soibam et al., 2023; Malakhov et al., 2023), and Kim and Park extending the analysis to varying flow regimes through a dedicated network design (Kim and Park, 2021). To extend segmentation performance beyond tightly controlled conditions, researchers have explored more advanced network architectures. Hessenkemper et al. (2022) compared U-Net (Ronneberger et al., 2015), StarDist (Schmidt et al., 2018), and Mask R-CNN (He et al., 2017), finding that a hybrid of U-Net and StarDist yielded the most robust results across variable scenarios. In parallel, Haas et al. (2020) introduced BubCNN, a composite model combining Faster R-CNN (Ren et al., 2015) with a shape regression module trained on over 100,000 annotated bubbles. Still, even with extensive training data, these models struggled with dense bubble clusters, elevated void fractions, and non-uniform lighting conditions, which are commonly encountered in bubbly datasets.

In recent years, transfer learning has emerged as a potential solution to this additional challenge. Cui et al. fine-tuned a COCO-pretrained Mask R-CNN on just 70 images, achieving accurate results up to 14.7% gas holdup (Cui et al., 2022). Homan and Deen used synthetic single-bubble masks to adapt a ResNet-50-based Mask R-CNN for use on modest hardware (Homan and Deen, 2024). Wang et al. explored SAM-assisted pipelines for real-time segmentation (Wang et al., 2025), and Xu et al. provided the first systematic evaluation of SAM's capabilities for this task (Xu et al., 2024), while Khojasteh et al. (2024) also reported promising SAM-based results on bubbly flows. Among these, SAM-based approaches stand out as particularly promising—despite current challenges with overlapping mask generation seen in SAM. This limitation is increasingly addressed by hierarchical SAM2 encoders, which aggregate multi-resolution features to enhance segmentation of

* Corresponding author.
E-mail addresses: S.Kuecuek@tudelft.nl (S. Küçük), C.DellaSantina@tudelft.nl (C.D. Santina), A.Laskari@tudelft.nl (A. Laskari).

https://doi.org/10.1016/j.ijmultiphaseflow.2025.105557
Received 30 September 2025; Received in revised form 11 November 2025; Accepted 23 November 2025
Available online 25 November 2025

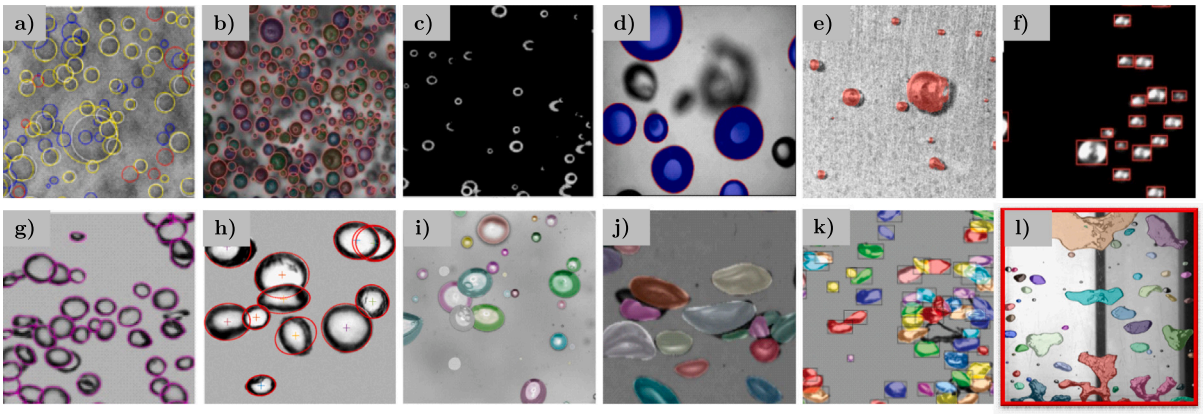

Fig. 1. Segmentation results from various deep learning-based approaches used in two-phase flow studies, ordered from round to increasingly deformable morphologies. (a) classical vs data-driven comparison by Ilonen et al. (2018), (b) transfer learning with COCO-pretrained Mask R-CNN by Cui et al. (2022), (c) hybrid RHT + Neural Network method by Serra et al. (2020), (d) Malakhov et al. (2023) under varying pressures (Malakhov et al., 2023), (e) CNN-based segmentation by Soibam et al. (2023), (f) real-time segmentation with SAM-assisted YOLO by Wang et al. (2025), (g) universal CNN model by Kim and Park (2021), (h) BubCNN model using Faster R-CNN and shape regression by Haas et al. (2020), (i) BubSAM model based on the Segment Anything Model (SAM) by Xu et al. (2024), (j) comparative study of U-Net, StarDist, and Mask R-CNN by Hessenkemper et al. (2022), (k) low-data transfer learning using ResNet-50 and synthetic images by Homan and Deen (2024), and (l) segmentation output from the model proposed in this study.

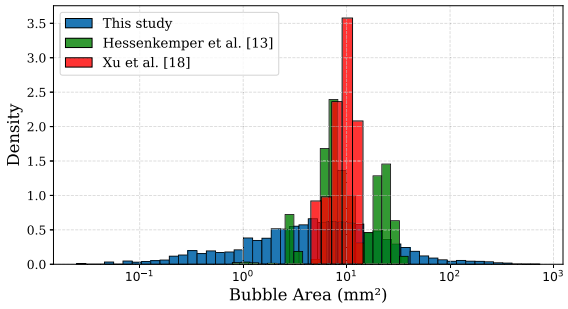

Fig. 2. Probability density function of the (wall-projected) area of all identified bubbles, including the current dataset (blue) compared with data from Hessenkemper et al. (2022) and Xu et al. (2024); broader distributions indicate increased multi-dispersity in air bubble size. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

overlapping or occluded structures, overcoming issues observed with the SAM. (Ravi et al., 2024; Xiong et al., 2024). Still, despite this substantial recent progress, a key limitation persists, as illustrated in Fig. 1: most studies focus on nearly spherical, isolated bubbles, often confined to narrow size ranges. This reduces variability and simplifies the learning task, but fails to capture the complex, deformable morphologies that characterize real-world multiphase flows.

In this work, we present an empirical investigation of bubble segmentation under complex, real-world conditions. We evaluate SAM 2.1 for the first time on a bubble segmentation task, benchmarking its ability to segment dense, irregular, and size-varying bubble structures.

Our dataset spans a size distribution several orders of magnitude wider than those used in prior work and includes bubbles ranging from perfectly spherical to highly non-convex and topologically complex shapes (Fig. 1 (l)). We further investigate how fine-tuning and data augmentation impact performance in these challenging settings. As a byproduct, we publicly release the labeled dataset used in our study, aiming to support future research in advancing bubble segmentation beyond the simplified regime of isolated, near-spherical bubbles.

2. Methodology

2.1. Dataset

The dataset used in this study originates from prior experimental work by one of the authors (Laskari, 2025), focused on air lubrication flows. This case was chosen for its practical relevance and the inherent complexity of the air phase topology, including not only discrete bubbles but also merged, elongated, and irregular air patches that deform continuously under turbulent flow conditions. The experiments were performed in a turbulent boundary layer flow over a flat plate fitted with a slot-type air injector. A Phantom 640-L high-speed camera, equipped with a 105 mm lens, was positioned above the plate to capture the air phase at 500 Hz. Illumination was provided by two LED panels ensuring high-contrast images suitable for detecting bubbles. The complete dataset contains several thousand images, recorded under varying flow conditions.

For this work, we selected 350 images corresponding to the bubbly flow regime and manually annotated each of them. To ensure temporal independence, we down-sampled time resolved sets to 5 Hz, minimizing the likelihood that the same bubble appears in multiple frames.

We generate masks that closely follow the outlines of the bubbles, rather than relying on bounding boxes. Note indeed that capturing detailed shape and contour information is essential for tasks such as deformation analysis and centroid estimation. Fig. 5 (a) illustrates a representative annotation from the dataset, for which the average gas hold-up is 19.5%.

2.2. Bubble categorization

As already discussed in the introduction, past research focuses on well-shaped regular bubbles with limited variation in size and morphology. In this subsection, we briefly report on the characteristics of the proposed dataset to support the claim of a substantial increase in complexity. In Fig. 2, we compare the estimated probability density function on a logarithmic scale of our dataset against the two recent studies in the literature for which we could find extensive size data (Hessenkemper et al., 2022; Xu et al., 2024). This dataset includes structures that are not only smaller but also significantly larger than those found in earlier studies. For example, the large end of the spectrum includes air patches, which are particularly relevant for drag reduction in air lubrication systems. Given the wide size

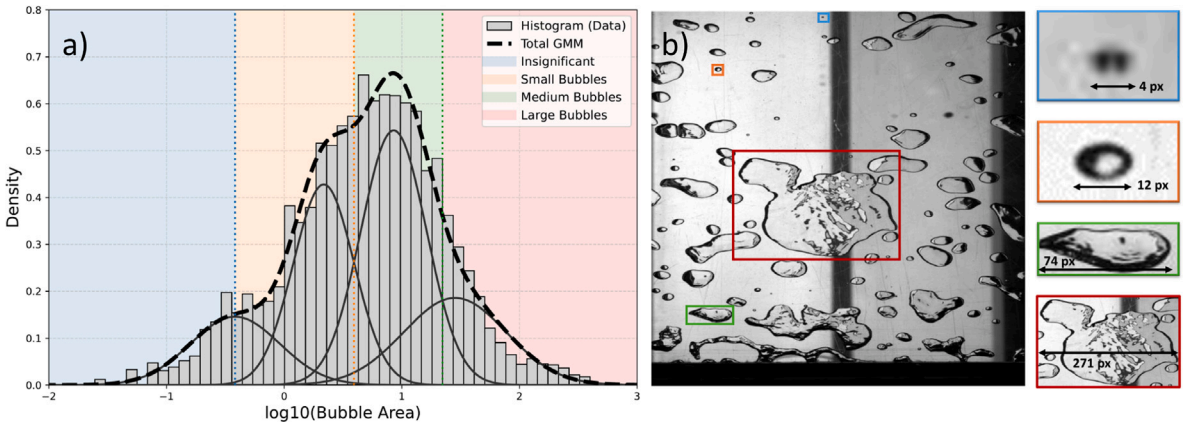

Fig. 3. (a) Gaussian Mixture Model (GMM) fit to the bubble area distribution, plotted in a semi-logarithmic scale. The GMM (dashed black) captures the multi-modal structure of the data, with its individual clusters shown as thin curves. The regions corresponding to small, medium, and large bubbles are color-coded in the background of the histogram as orange, green, and red, respectively, with thresholds set at the intersection points of the Gaussians. Values below the mean of the smallest (insignificant) cluster were excluded as noise and color-coded with blue. (b) Example bubbles corresponding to each GMM cluster are shown, with bounding boxes color-coded according to their respective cluster category. The size of each bubble in pixels is also shown, giving a clear idea of the differences between clusters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

variability, we performed a statistical analysis of the bubble area distribution generally well described by a lognormal (see Mouza et al., 2005 and Jacob et al., 2010, among many others), which when plotted in a logarithmic scale, is expected to follow a near-Gaussian trend. We then fitted a Gaussian Mixture Model (GMM), selecting the optimal number of components using the Bayesian Information Criterion. As shown in Fig. 3, (a) the Gaussian Mixture Model (GMM, dashed black line) provided an accurate representation of the data. Later in our proposed data augmentation strategy, and the discussion of the results, we interpreted the individual components (indicated with different colors and demarcated by the vertical dashed lines) of the GMM as meaningful clusters. The smallest cluster was the most sensitive to resolution issues and noise, so we excluded values below its mean (bubble projected areas < 0.5 mm^2 and classified as *insignificant*). In the rest of the paper, we refer to the remaining three clusters as small, medium, and large bubbles, with thresholds set at the intersection points of the three Gaussians. Specifically, bubbles with areas between about 0.5 and 5 mm^2 were classified as *small*, those between 5 and 30 mm^2 as medium, and those larger than 30 mm^2 as large. In the histogram, the regions corresponding to insignificant, small, medium, and large bubbles are color-coded in the background as blue, orange, green, and red, respectively. Example bubbles from each category, along with their size (in pixels), are shown in Fig. 3(b). These include bubbles with areas ranging from below ≈ 1 mm^2 up to 1000 mm^2. This wide range of bubble size, and especially the presence of very large ones — for which surface tension stops having a dominant effect with resulting shapes diverging considerably from spherical — leads also to a considerable range of bubble shapes, increasing the complexity of the data.

2.3. Data augmentation strategy

During preliminary evaluations of SAM 2.1 on the proposed segmentation, we quickly realized that large air patches were frequently observed and well detected. At the same time, medium and small bubbles were less accurately identified (quantitative assessment is available in the Results section). Thus, we decided to focus our data augmentation efforts on improving the model's performance for these smaller structures. During pre-processing, we auto-oriented the images and cropped them to retain only the region along the flow direction, corresponding to the upper half, where small and medium-sized bubbles are more concentrated. The images were resized to 640×640 pixels to ensure consistency. Offline augmentations were performed at the mask level, directly modifying the segmentation masks to simulate realistic variations in the data. These included adding random noise affecting up to 0.1% of pixels and applying shear transformations of up to ± 10 degrees horizontally and vertically around the bubble boundaries.

In addition to these mask-level augmentations, standard image-level online augmentations were also used during training and runtime. These included random flipping and color adjustments, as specified in the SAM 2.1 training configuration file. This hybrid augmentation strategy increased the diversity of training data and enhanced the generalization capability.

2.4. Metrics

We define Precision and Recall as $\mathcal{P} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, $\mathcal{R} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, where TP and FP are true and false positives, and TN and FN are true and false negatives. Precision measures how many of the model's positive predictions are correct, while Recall indicates how many of the actual positive cases the model correctly identifies. F1 score combines precision and recall into a single harmonic mean, enabling balanced evaluation. $\mathrm{F1} = 2\mathcal{P}\mathcal{R}/(\mathcal{P}+\mathcal{R})$. This metric is commonly used to balance the trade-off between precision and recall.

To evaluate segmentation quality, we report Intersection over Union (IoU) and Dice similarity, which quantify mask accuracy and boundary alignment. Dice, being more sensitive to overlap, is particularly informative for irregular or small structures. More precisely, IoU is defined as $(A \cap B)/(A \cup B)$, while Dice is $2(A \cap B)/(A + B)$, where A is the predicted mask and B is the ground truth mask.

2.5. Fine-tuning strategy

The fine-tuning process was carried out using the SAM 2.1 training framework, which comes with built-in tools for model development and adjustment. By offering ready-to-use settings for training steps, model configuration, and logging, it helps streamline the overall process. The framework also manages technical aspects, such as GPU usage, saving training checkpoints, and utilizing mixed-precision to accelerate computation. Thanks to its support for multi-GPU training with PyTorch's DistributedDataParallel (DDP), it runs efficiently on both single systems and larger computing setups. More details on the training setup and code can be found in the official SAM 2.1 documentation (Ravi et al., 2024).

The training configuration was defined in a separate YAML file, allowing precise control over which model components are trainable

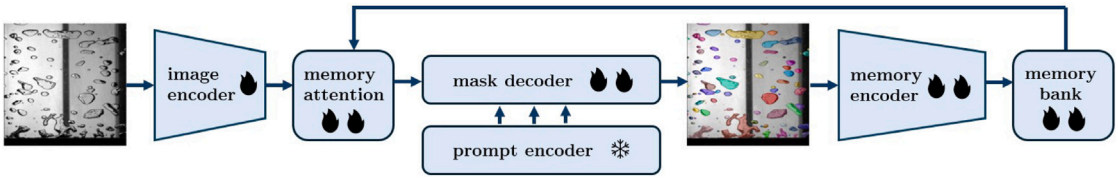

Fig. 4. Architecture of the SAM 2.1 model used in our experiments. Modules marked with a snowflake were kept frozen, while those marked with a flame were fine-tuned. A single flame indicates partial tuning (e.g., only the final block of the image encoder), while two flames denote fully trainable modules such as the memory encoder and mask decoder.

and how training parameters are set. In this setup, the image encoder was partially fine-tuned using a lower learning rate of 3e−6 to retain pretrained features; specifically, only the trunk layers were updated, while the neck layers and embedding layers remained frozen to preserve their pretrained weights. In contrast, the mask decoder, memory attention, and memory encoder modules were fully trainable and used a higher learning rate of 5e−6 to enable faster adaptation to the object segmentation task. This two-level learning rate setup keeps the general features steady while letting the task-specific parts learn faster. We also used a cosine annealing schedule to gradually lower the learning rates, which helped training be more stable and the model generalize better. The model was trained for 150 epochs using the AdamW optimizer, chosen for its better regularization and more effective handling of weight decay compared to standard Adam. The batch size was set to the maximum that fit in GPU memory — 3 per GPU — and automatic mixed precision was enabled to reduce memory usage and speed up training. Regular augmentations like affine transforms, flipping, and color jitter improved generalization. The final learning rates, optimizer, and number of epochs were selected through preliminary trials on a held-out 20% validation subset. Training and validation loss curves were monitored to balance convergence speed and overfitting. The chosen setup corresponded to the point where validation loss stabilized while training loss decreased slowly, indicating efficient learning and stable generalization. Optimizer settings, training duration, and logging intervals were all managed through the same config file to ensure reproducibility. The overall SAM 2.1 architecture is shown in Fig. 4. More details and the training code are available on the project's GitHub repository.

A multi-part loss combining cross-entropy, Dice, IoU, and classification losses was used, with extra weight on spatial overlap terms to handle irregular object boundaries better. Cross-entropy, being differentiable, allows effective training by reducing prediction errors, which indirectly improves precision and recall since these metrics are not directly optimized during training. The relative weighting of loss terms determines the model's focus. Emphasizing the Cross-entropy component encourages balanced detection of small and large bubbles, whereas prioritizing the IoU term favors accurate segmentation of larger structures. In this study, we assigned slightly higher weight to IoU to ensure robust detection of dominant bubbles — most relevant for interpreting air-lubrication flows — while accepting that some small bubbles might occasionally be missed. This trade-off was deliberate, as most conventional and learning-based methods already perform well for small, nearly spherical bubbles under simple geometric assumptions, whereas accurately segmenting large, irregular, and coalesced structures remains the main open challenge. As a result, the model yielded more stable predictions in challenging flow regimes, particularly where bubble deformation, merging, or breakup led to ambiguous boundaries.

3. Results

3.1. Baseline evaluation and model selection

To establish a reference point, the baseline performances of both the original SAM and SAM 2.1 models were first examined. SAM achieved

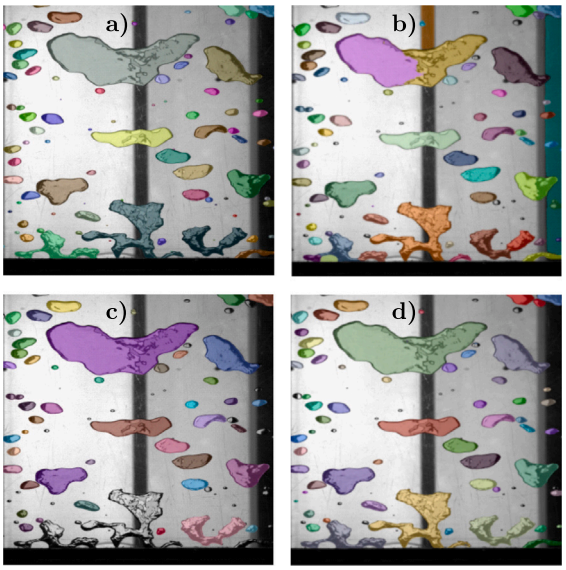

Fig. 5. Visual comparison of segmentation results. The air injector is located at the bottom of the images (shaded in black), and the flow is from bottom to top. (a) Manually annotated ground truth masks used as reference. (b) Output of the base Segment Anything Model (SAM). (c) Output of the base (i.e., non-fine-tuned) SAM 2.1 model. (d) Segmentation results from the fine-tuned SAM 2.1 model developed in this study, trained on 240 images using the Augmentation 2 strategy.

an overall F1 score of 0.705, while the non-fine-tuned SAM 2.1 base model reached a slightly higher score of 0.720. Although SAM's performance may initially seem competitive, a closer examination reveals certain limitations, particularly in the detection of large bubbles. While the model achieved a very high recall of 0.966 for large bubbles, its precision was only 0.479. This indicates that SAM often generated more than one mask for a single large bubble, leading to over-segmentation. Such errors can be critical in applications where quantifying the accurate number and size of bubbles is essential. In contrast, the SAM 2.1 base model provided a more balanced performance for large bubbles, with a precision of 0.934 and a recall of 0.903. This improved trade-off between precision and recall highlights SAM 2.1 as the preferred candidate for the fine-tuning process in this study.

This quantitative observation is supported by qualitative visual inspection of the masks produced by both models, superimposed on the raw images and compared against the ground truth annotations. As shown in Fig. 5, which reports the segmentation results for a representative frame, the baseline SAM model (Fig. 5b) successfully detects most of the bubbles and air patch regions. However, it also generates several spurious masks over background areas unrelated to the air phase—see, for instance, the shaded region on the right side of Fig. 5b. Additionally, the model often splits individual objects into multiple adjacent masks (see the top of Fig. 5b) or creates overlapping ones for the same object (not easily visually distinguishable in the figure). These observations indicate that, while SAM exhibits high sensitivity, it lacks

precision in delineating object boundaries and frequently produces masks that are irrelevant or redundant. In contrast, the SAM 2.1 base model (Fig. 5c) detects many air pockets while avoiding the generation of multiple masks for the same object, resulting in more coherent and consistent segmentations. Some irregularly shaped patches with ambiguous boundaries are missed, likely due to the model's conservative prediction strategy, which only assigns masks when confident that a meaningful object is present. In this context, such caution is preferable to the over-segmentation seen in the earlier SAM version. Importantly, this behavior suggests that with appropriate fine-tuning on data containing sufficient air-phase exemplars, SAM 2.1 could achieve robust and reliable segmentation performance.

Based on these quantitative and qualitative findings, we selected the SAM 2.1 model as the foundation for fine-tuning in this study. Fig. 5d presents qualitative results from the fine-tuned model, trained on a dataset of 240 images (comprising 100 manually annotated, 80 for training 20 for validation, and 160 augmented samples), alongside the baseline comparisons. The segmentation performance shows clear improvement: the model accurately captures the boundaries of air pockets — particularly medium and large ones — while only occasionally missing smaller bubbles. Detailed quantitative results and a description of the dataset construction are provided in the following section.

3.2. Fine-tuning and data augmentation performance

To systematically evaluate the performance of the fine-tuned SAM 2.1 model and quantify the impact of data augmentation, we report F1 and Dice scores on a fixed validation set across three training sets of increasing size, each corresponding to a different augmentation strategy (Fig. 6, left). The first set includes only manually labeled real images (No Augmentation), the second combines 50% real and 50% offline augmented images (Augmentation Rank 1), and the third uses one-third real and two-thirds augmented images (Augmentation Rank 2). This setup enables an empirical assessment of the trade-off between annotation effort (required for real images) and computational overhead (minimal for augmentations), and allows us to identify an optimal balance. For reference, we also compare these results with those from the baseline SAM 2.1 model (i.e., without fine-tuning, corresponding to a training set size of zero in Fig. 6). Since our application involves segmenting images with highly variable bubble sizes, we report metrics both globally — across all detected bubbles (Fig. 6 a4) — and disaggregated by size category: small (Fig. 6 a1), medium (Fig. 6 a2), and large (Fig. 6 a3).

When looking at the overall metrics, for the smallest training set size tested (50 non-augmented images), the fine-tuned model already exhibits a 12% and 11% increase from the baseline case for Dice and F1 scores, respectively (Fig. 6 a4). As the training set size increases, performance keeps increasing monotonically (for all three augmentation strategies and both metrics) and then plateaus, beyond a set size of 240 images. A training dataset larger than that, irrespective of augmentation strategy, does not provide a meaningful performance improvement; thus for the rest of this section we will focus on results for training datasets of this size. This choice was also reflected in the visual result comparison above, where the output from the SAM 2.1 model, fine-tuned using 240 images with an augmentation of Rank 2 was included (Fig. 5d).

When comparing augmentation strategies, peak performance is very similar across all three, with F1 scores ranging from 0.815 to 0.837 and Dice scores reaching 0.929 to 0.937, steadily increasing from no augmentation to Augmentation Ranks 1 and 2. These results confirm that the fine-tuned model performs remarkably well under all tested configurations—achieving high-quality segmentation even when trained with limited manually labeled data. Two main conclusions follow from these observations. First, a training set composed of only 80 manually annotated images, complemented by augmented data, achieves performance on par with — or even slightly better than — that of a model trained on 240 real images. This demonstrates that data augmentation can significantly reduce annotation burden without compromising accuracy. Second, the marginal differences between Rank 1 and Rank 2 augmentation strategies suggest that the benefit of additional offline augmentation saturates, with limited gains beyond a certain point.

When the results are examined across different bubble size categories, performance is seen to vary significantly. Medium and large bubbles are segmented with high accuracy, exhibiting F1 scores between $0.95 - 0.965$ and Dice between $0.935 - 0.971$ (higher scores for the large bubbles). The performance metrics associated with large bubble sizes also differ slightly from the global ones discussed above, in terms of their variation with increasing training set size: performance does not monotonically increase but rather oscillates, especially when augmented images are used. This difference can be attributed to the fact that augmented data were created using only the top part of the original images, where mostly small and medium sized bubbles are present, thus somewhat limiting the benefits of augmentation for large bubbles, particularly for the F1 scores. Regardless, overall accuracy for these bubbles is the highest and is also seen to plateau around a similar training dataset size. In contrast, for small bubbles, although fine-tuning of the base model leads to almost doubling of both F1 and Dice scores, these still remain significantly lower, with values around 0.577 and 0.503, respectively. These results are chiefly due to our training choices, where lower weights were used for loss functions associated with small bubbles, given their relatively lower importance in the application of air lubrication. This trend is reinforced by the weighting between the F1 and IoU losses. While the F1 loss treats all detections equally, the IoU loss imposes stronger penalties for larger missed regions, slightly biasing the model toward better segmentation of large bubbles that are more relevant to flow analysis.

Finally, seeing that both illumination and bubble size distribution across the original images was non-uniform (Fig. 5), we also assess here whether there is any resulting spatial variation in model performance across the image plane. When looking at the recall values of the baseline SAM 2.1 model (Fig. 6b, bottom), we can see that there is indeed a spatial inhomogeneity present, with lower recall values close to the air injection, where larger, irregularly shaped air patches are present, and also on the right side, where illumination is insufficient. In contrast, the fine-tuned model (augmentation of Rank 2, for 240 training images) shows higher recall values overall, as expected, and no spatial dependence in performance (Fig. 6b, top) with consistent segmentation across all regions, highlighting another gain from the fine-tuning process.

3.3. Model comparison

Following the observed improvements after fine-tuning, we compared SAM 2.1 with a conventional segmentation network. For this purpose, a Mask R-CNN model with a ResNet-50 backbone, pre-trained on COCO, was trained on the same 100-image dataset as SAM 2.1, with 80% of the images used for training and the remaining 20% for validation. Table 1 presents the results of this comparison, showing the performance of the Base SAM 2.1 model alongside the fine-tuned Mask R-CNN and fine-tuned SAM 2.1. Fine-tuned models were trained on the same 100-image dataset, allowing for a fair and direct evaluation across overall, small, medium, and large bubble categories. For small, nearly spherical bubbles, Mask R-CNN performs slightly better, capturing these simple structures with reasonable accuracy. However, when bubble shapes deviate from the ideal sphere, Mask R-CNN often fails to detect the full objects. Its strong edge sensitivity causes it to assign masks to every sharp intensity transition, including internal reflections and illumination artifacts within bubbles. This occasionally helps in identifying small, high-contrast bubbles, but leads to severe over-segmentation and fragmented masks for larger or irregular shapes.

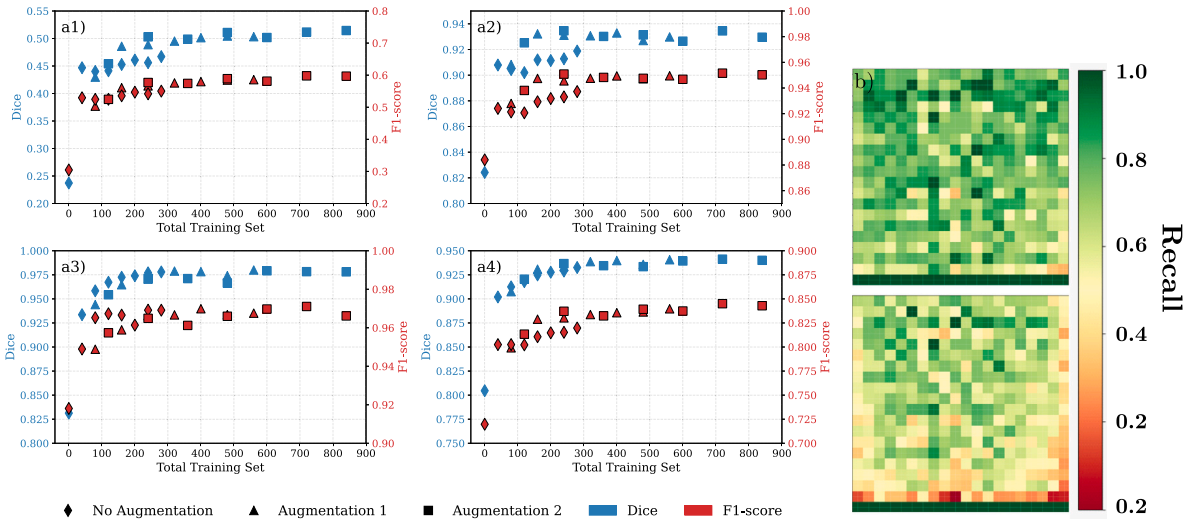

Fig. 6. Segmentation performance of fine-tuned SAM2.1. (a) F1 and Dice scores for models trained on datasets of different sizes and with various augmentation strategies, shown for small (a1), medium (a2), and large (a3) bubbles (see Section 2.2), and for all identified bubbles (a4). Evaluations are performed across three augmentation strategies: No Augmentation, Augmentation 1 (50% real and 50% augmented data), and Augmentation 2 (33% real and 66% augmented data). (b) Spatial variation of model performance (based on recall values). Top: Fine-tuned SAM 2.1 using 240 training images with Augmentation 2 strategy. Bottom: Baseline SAM 2.1.

In contrast, SAM 2.1 leverages global visual context and hierarchical features to interpret the full extent of deformable and coalesced bubbles, maintaining mask coherence even under non-uniform illumination. As a result, SAM 2.1 consistently outperforms Mask R-CNN for medium and large bubbles, offering more physically meaningful and topologically consistent segmentation, while its performance on small, simple bubbles remains comparable. This difference is due to the training strategy, which prioritized larger structures, leading to lower F1-score for small bubbles.

The above analysis highlights the capabilities of a fine-tuned SAM 2.1 model in segmenting bubbles with a wide range of sizes and shapes, even in regions of insufficient illumination (see Fig. 6b). More broadly, applicability of this model in other experimental setups and flow geometries, will depend only implicitly on variations of injector type, void fraction, or optical conditions: it would be the resulting bubble size distribution and the relative prevalence of small versus large bubbles, which could potentially impact performance. Yet, given the strong baseline performance and generalization capabilities of SAM 2.1, we expect that with tailored adjustments to the fine-tuning strategy for specific purposes researchers would be able to achieve robust segmentation in a wide range of scenarios.

Table 1
Comparison of segmentation performance among the Base SAM 2.1 model, fine-tuned Mask R-CNN, and fine-tuned SAM 2.1 across different bubble size ranges. The Base model represents the unmodified pre-trained variant, while both Mask R-CNN and SAM 2.1 were fine-tuned on the same 100-image dataset (80 images for training and 20 for validation) to ensure a fair comparison.

Metric	Base SAM 2.1	Fine-Tuned Mask R-CNN	Fine-Tuned SAM 2.1
Small bubbles			
F1-score	0.3045	**0.7752**	0.5246
Dice-score	0.2372	**0.7249**	0.4401
Medium bubbles			
F1-score	0.8840	0.6667	**0.9213**
Dice-score	0.8242	0.5809	**0.9044**
Large bubbles			
F1-score	0.9181	0.7325	**0.9652**
Dice-score	0.8313	0.7442	**0.9584**
Overall			
F1-score	0.7197	0.7612	**0.8025**
Dice-score	0.8046	0.7163	**0.9125**

4. Conclusion

This study demonstrated that high-quality bubble segmentation, across a wide range of sizes and shapes, can be achieved with minimal annotated data. Fine-tuning SAM 2.1 on as few as 100 labeled images (including both the training and validation sets) resulted in substantial gains, particularly for medium and large bubbles, with F1 and Dice scores approaching 0.95. While small bubbles remain more challenging, this trade-off reflects an intentional bias in training priorities, deliberately favoring stable detection of the most physically relevant structures rather than indicating a fundamental limitation. This observation aligns with previous findings: state-of-the-art geometric models are sufficient for small, near-spherical bubbles, while segmenting larger, irregular structures remained an open challenge. Future work could focus on ensemble approaches, combining classic techniques for small bubbles, and more advanced DL-based segmentation methods for non-trivial bubbles. Finally, compared to existing approaches that require extensive datasets yet struggle with generalization, our method offers a data-efficient alternative that remains robust in complex, real-world flow conditions. To support broader adoption, we release both our labeled dataset and fine-tuning pipeline—aiming to make accurate, low-effort segmentation accessible to the multiphase flow community.

While not universally applicable to all two-phase flows, our results demonstrate that accurate segmentation is achievable under challenging conditions, and that the SAM 2.1 workflow can be readily adapted to new setups with minimal labeled data. Future work will focus on applying Fine-tuned SAM 2.1 as a tool in large-scale, multi-phase flow studies, as well as investigating bubble tracking.

CRediT authorship contribution statement

Semanur Küçük: Writing – original draft, Validation, Project administration, Methodology. **Cosimo Della Santina:** Writing – review & editing, Supervision, Project administration. **Angeliki Laskari:** Writing – review & editing, Supervision, Project administration, Data curation.

AI Declaration

AI-assisted tools were used solely to improve the readability and language of the manuscript. All scientific content, analysis, and interpretations remain the original work of the authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

To ensure full reproducibility, both the dataset and the code used in this study are publicly available. The dataset can be accessed on Kaggle (https://www.kaggle.com/datasets/semanurkk/bubblyflow), including all images and annotations split into 80% training and 20% validation sets. Images are provided in `.jpg` format, and annotations in `.json` files following the SAM 2 structure for direct use in model training and evaluation. The accompanying code is available on GitHub (https://github.com/Semanur97/BubblyFlow.git).

References

Cui, Y., Li, C., Zhang, W., Ning, X., Shi, X., Gao, J., Lan, X., 2022. A deep learning-based image processing method for bubble detection, segmentation, and shape reconstruction in high gas holdup sub-millimeter bubbly flows. Chem. Eng. J. 449, 137859.

Haas, T., Schubert, C., Eickhoff, M., Pfeifer, H., 2020. Bubcnn: Bubble detection using faster rcnn and shape regression network. Chem. Eng. Sci. 216, 115467.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

Hessenkemper, H., Starke, S., Atassi, Y., Ziegenhein, T., Lucas, D., 2022. Bubble identification from images with machine learning methods. Int. J. Multiph. Flow 155, 104169.

Homan, T.A., Deen, N.G., 2024. Deep learning bubble segmentation on a shoestring. Ind. Eng. Chem. Res. 63 (17), 7800–7806.

Ilonen, J., Juránek, R., Eerola, T., Lensu, L., Dubská, M., Zemčík, P., Kälviäinen, H., 2018. Comparison of bubble detectors and size distribution estimators. Pattern Recognit. Lett. 101, 60–66.

Jacob, B., Oliveri, A., Miozzi, M., Campana, E.F., Piva, R., 2010. Drag reduction by microbubbles in a turbulent boundary layer. Phys. Fluids 22, 115104.

Khojasteh, A.R., van de Water, W., Westerweel, J., 2024. Practical object and flow structure segmentation using artificial intelligence. Exp. Fluids 65 (8), 119.

Kim, Y., Park, H., 2021. Deep learning-based automated and universal bubble detection and mask extraction in complex two-phase flows. Sci. Rep. 11 (1), 8940.

Laskari, A., 2025. Effects of liquid turbulent boundary layer spanwise organisation on air lubrication. Int. J. Multiph. Flow (in press).

Malakhov, I., Seredkin, A., Chernyavskiy, A., Serdyukov, V., Mullyadzanov, R., Surtaev, A., 2023. Deep learning segmentation to analyze bubble dynamics and heat transfer during boiling at various pressures. Int. J. Multiph. Flow 162, 104402.

Mouza, A.A., Dalakoglou, G.K., Paras, S.V., 2005. Effect of liquid properties on the performance of bubble column reactors with fine pore spargers. Chem. Eng. Sci. 60, 1465–1475.

Ni, R., 2024. Deformation and breakup of bubbles and drops in turbulence. Annu. Rev. Fluid Mech. 56 (1), 319–347.

Qin, S., Chu, N., Yao, Y., Liu, J., Huang, B., Wu, D., 2017. Stream-wise distribution of skin-friction drag reduction on a flat plate with bubble injection. Phys. Fluids 29 (3).

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al., 2024. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Schmidt, U., Weigert, M., Broaddus, C., Myers, G., 2018. Cell detection with star-convex polygons. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11. Springer, pp. 265–273.

Serra, P., Masotti, P., Rocha, M., de Andrade, D., Torres, W., de Mesquita, R., 2020. Two-phase flow void fraction estimation based on bubble image segmentation using randomized hough transform with neural network (rhtn). Prog. Nucl. Energy 118, 103133.

Soibam, J., Scheiff, V., Aslanidou, I., Kyprianidis, K., Fdhila, R., 2023. Application of deep learning for segmentation of bubble dynamics in subcooled boiling. Int. J. Multiph. Flow 169, 104589.

Tanaka, T., Oishi, Y., Park, H.J., Tasaka, Y., Murai, Y., Kawakita, C., 2023. Downstream persistence of frictional drag reduction with repetitive bubble injection. Ocean Eng. 272, 113807.

Wang, B., Lv, H., Wang, X., Hao, M., Kirk, D., Guay, D., Thorpe, S., Ruan, Z., 2025. Quantifying bubble-induced diffusion resistance through real-time sam-assisted yolo high density bubble detection algorithm. Chem. Eng. J. 512, 162422.

Wang, X.-y., Su, H.-c., Li, S.-w., Wu, G.-h., Zheng, X.-x., Duan, Y.-x., Zhang, Y.-n., 2023. Experimental research of the cavitation bubble dynamics during the second oscillation period near a spherical particle. J. Hydrodyn. 35 (4), 700–711.

Xiong, X., Wu, Z., Tan, S., Li, W., Tang, F., Chen, Y., Li, S., Ma, J., Li, G., 2024. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. arXiv preprint arXiv:2408.08870.

Xu, H., Feng, X., Pu, Y., Wang, X., Huang, D., Zhang, W., Duan, X., Chen, J., Yang, C., 2024. Bubsam: Bubble segmentation and shape reconstruction based on segment anything model of bubbly flow. AIChE J. 70 (12), e18570.