AI Thermal Fluids 5 (2026) 100028

Contents lists available at ScienceDirect

AI Thermal Fluids

journal homepage: www.sciencedirect.com/journal/ai-thermal-fluids

Multi-modality fusion: Illuminating the complexity of two-phase flow

Xiaojing Zhang [a], Sanghyeon Chang [a], Youngjoon Suh [a,b], Yoonjin Won [a,c,*]

[a] Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, United States of America
[b] Department of Mechanical and Industrial Engineering, University of Illinois, Chicago, IL, United States of America
[c] Department of Electrical and Computer Engineering, University of California, Irvine, CA, United States of America

ARTICLE INFO

Keywords:
Multi-modality
Pre-fusion
Early-fusion
Segmentation
Two-phase flows

ABSTRACT

Bubble dynamic analysis is crucial for modeling and optimizing phase change heat transfer. Recent advances in computer vision models have been widely applied to automatically extract bubble behavior parameters related to their performances. However, current segmentation models heavily rely on large datasets to achieve high accuracy, imposing a significant human workload. To overcome this challenge, we propose a novel data-fusion approach that enriches input information with dynamic features of bubble motion. Our methodology incorporates multi-modal generation techniques to transform conventional high-speed images into complementary modalities, and then fuse them for enhanced representation. This framework enables precise segmentation prediction using small datasets and without any pre-trained models. The resulting bubble statistics provide a quantitative foundation for advancing boiling heat transfer analysis and elucidating interfacial transport mechanisms. Overall, our multi-modal segmentation framework offers an automated, data-efficient, accurate, and versatile method for bubble behavior quantification in diverse two-phase flow systems.

1. Introduction

Bubble behavior serves as a phenomenological representation of how heat transfer occurs and performs during phase change processes like boiling and condensation. Over the past decades, extensive efforts have been devoted to developing empirical boiling correlations that capture various aspects of bubble dynamics, from bubble nucleation to departure. The onset of nucleate boiling (ONB) is characterized by early-stage bubble dynamics, including nucleation site density and the initial growth and departure of isolated bubbles [1]. In contrast, critical heat flux (CHF) predictions rely on bubble features, such as departure diameter, departure frequency, and nucleation site density [2]. Additional indicators, including void fraction and wetting front area, have been shown to enhance correlations for heat transfer coefficients during flow boiling [3]. A prominent example includes the Rensselaer Polytechnic Institute (RPI) [4] model, which links bubble dynamics to heat transfer coefficients and critical heat flux predictions. Nevertheless, reliance on manual image analysis continues to impose significant workload demands and introduces the risk of human error.

The recent advances in computer vision techniques show the potential to automate and improve the accuracy of bubble dynamics analysis. Among these techniques, instance segmentation offers a powerful tool to isolate and quantify individual objects, i.e., bubbles within complex multiphase flow systems. Segmentation models, such as mask region-based convolutional neural network (Mask R-CNN) [5], U-Net [6], and you only look once (YOLO)-based models [7], demonstrate strong performance across diverse image analysis tasks.

However, segmentation models often struggle under boiling conditions, where rapid bubble motions, uneven illumination, and extensive overlap are present. For instance, a Mask R-CNN trained across varied two-phase flow conditions reported significant drops in segmentation accuracy when lighting conditions changed or bubbles overlapped extensively [8]. Similarly, a U-Net model fine-tuned on high-speed subcooled boiling sequences achieved only moderate segmentation quality when bubbles coalesced or shuddered in low-contrast frames [9]. These limitations highlight the challenges posed by real boiling environments, where the inherent complexity of two-phase flows makes segmentation particularly difficult.

Among existing approaches, YOLOv8 shows strong potential to address these challenges through its anchor-free, decoupled head architecture, which enables precise instance segmentation of small and fast-moving bubbles—an essential capability for analyzing complex two-phase flows. For example, prior studies have demonstrated that YOLOv8 combined with Deep SORT, achieved higher detection accuracy and significantly faster inference speed than Mask R-CNN [10].

* Correspondence to: Department of Mechanical and Aerospace Engineering, University of California, Irvine, 4200 Engineering Gateway, Irvine, CA 92617, United States.
E-mail address: won@uci.edu (Y. Won).

https://doi.org/10.1016/j.aitf.2026.100028
Received 3 October 2025; Received in revised form 18 December 2025; Accepted 16 January 2026

3050-5852/© 2026 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

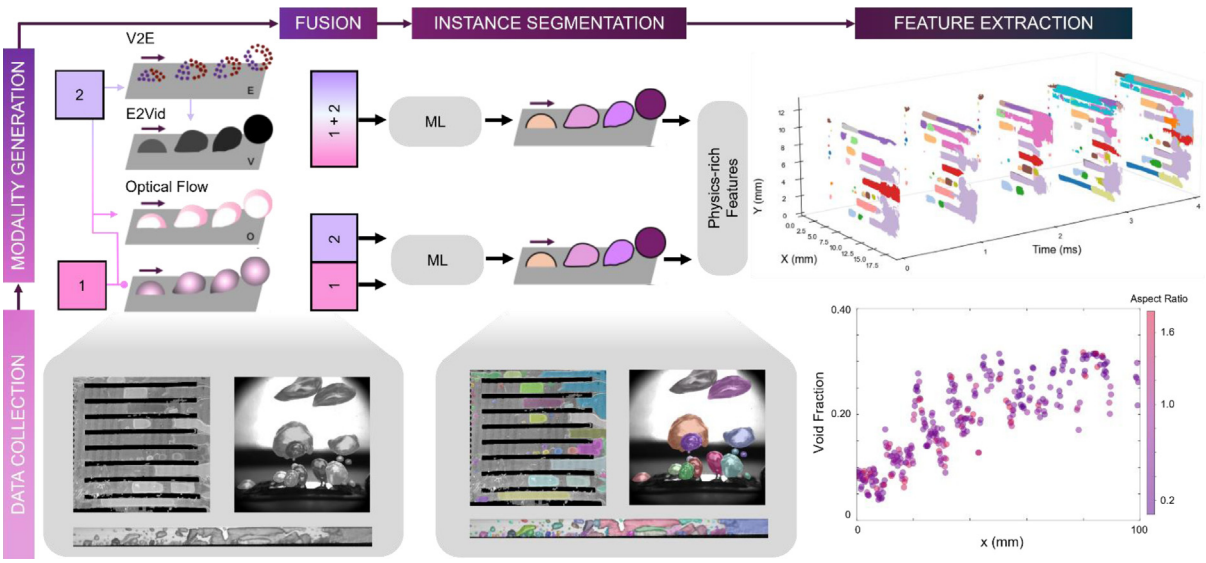

Fig. 1. A Framework Integrating Multi-Modal and Data Fusion. Data collection: high-speed images (1,024 ×1,024 pixels) from microflow boiling, flow boiling, and pool boiling experiments. **Modality generation**: collected high-speed images undergo multi-modal processing procedures including V2E (video-to-events), E2Vid (events-to-video reconstruction), and GMFlow (global matching flow), resulting in events (E) (red dots mean event on while purple dots mean event off), reconstructed video (V), and optical flow (O) modalities. **Fusion**: high-speed images and generated modalities are fused via pre-fusion and early-fusion. **Instance segmentation**: training enables models to identify bubbles and output their instance-level masks. **Feature extraction**: important bubble statistics like local void fraction and aspect ratio can be obtained from predicted bubble masks.

Another example on biological dataset, YOLOv8 outperformed Mask R-CNN in instance segmentation accuracy and dimensional measurement precision, particularly under low-quality imaging conditions [11].

Optimizing segmentation models for two-phase flows can be pursued through two main strategies. The *first* strategy is to increase the number of training images, thereby exposing the model to a wide range of bubble appearances for high accuracy [12–14]. However, supervised learning algorithms [15–17] used in segmentation model connects large dataset (thousands of images) with large human annotation effort. The alternative and promising strategy is to leverage different types modalities that can provide complementary information about the same objects and thereby enrich the input space [18]. Such modalities can arise from different imaging equipment (e.g., tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) in medicine [19]). Using multiple modalities helps the segmentation model better distinguish objects under challenging conditions [20–22].

The aforementioned methodologies highlight the potential of multi-modal data to enhance instance segmentation accuracy. Nevertheless, notable research gaps remain, as most prior work has focused on only a few tasks. The fusion of multi-modal has been limited to early- and late-fusion strategies, with little exploration of prefusion or alternative approaches. In addition, most existing research has emphasized large datasets (reach up to 95,000 images), overlooking the potential of multi-modal approaches on smaller datasets. Moreover, the majority of studies has focused on object detection rather than instance segmentation, making it difficult to transfer conclusions and model architectures to segmentation applications due to inherent differences in shape and other characteristics between two tasks.

To address these challenges, we propose a framework that integrates multi-modal datasets and their fusion, as illustrated in Fig. 1. The framework begins with the data collection of high-speed images from three different experimental setups which are then converted into multiple modalities – events (E), reconstructed video (V), and optical images (O) – using video-to-events (V2E), events-to-video reconstruction (E2Vid), and global matching flow (GMFlow), respectively. These modalities are subsequently processed through the proposed prefusion or early-fusion and their corresponding segmentation models. Finally, the fusion-aided segmentation outputs enable the extraction of quantitative bubble statistics providing the foundation for subsequent heat transfer analysis.

2. Data collection using high-speed cameras

The images in the dataset are captured using a high-speed camera at 2,000 fps with a light emitting diode (LED) from our (i) microflow boiling, (ii) flow boiling, and (iii) pool boiling experiments, as shown in Fig. 2. More information can be seen in *Supplementary Information Section* 1. In microflow boiling, top view is captured with the help of a mirror, emphasizing lateral bubble motion. In flow boiling and pool boiling, the camera captures a side view of the bubbles, providing information on their vertical growth, elongation, and detachment. All visualization measurements are taken at consistent focal lengths and light intensities for each setup to maintain fixed pixel-to-length ratios.

3. Multi-modality generation

Conventional high-speed imaging produces large data volumes because every pixel is recorded, including static backgrounds in addition to the moving bubbles. Capturing background structures also introduces irrelevant details that can obscure bubble boundaries and complicate segmentation. Furthermore, even with microsecond-scale exposures, limited light conditions or rapid bubble dynamics may still lead to motion blur. High-speed images are also sensitive to optical variations, which further degrade boundary clarity and lower the reliability of segmentation models. Thus, large training datasets are typically required for the model to learn robust bubble features.

These limitations can be mitigated by introducing alternative image modalities. The generated modalities include event-based representations (E), reconstructed video from event data (V), and optical flows (O). Implementation details and mechanism for all transformations are provided in *Supplementary Information Section* 2.

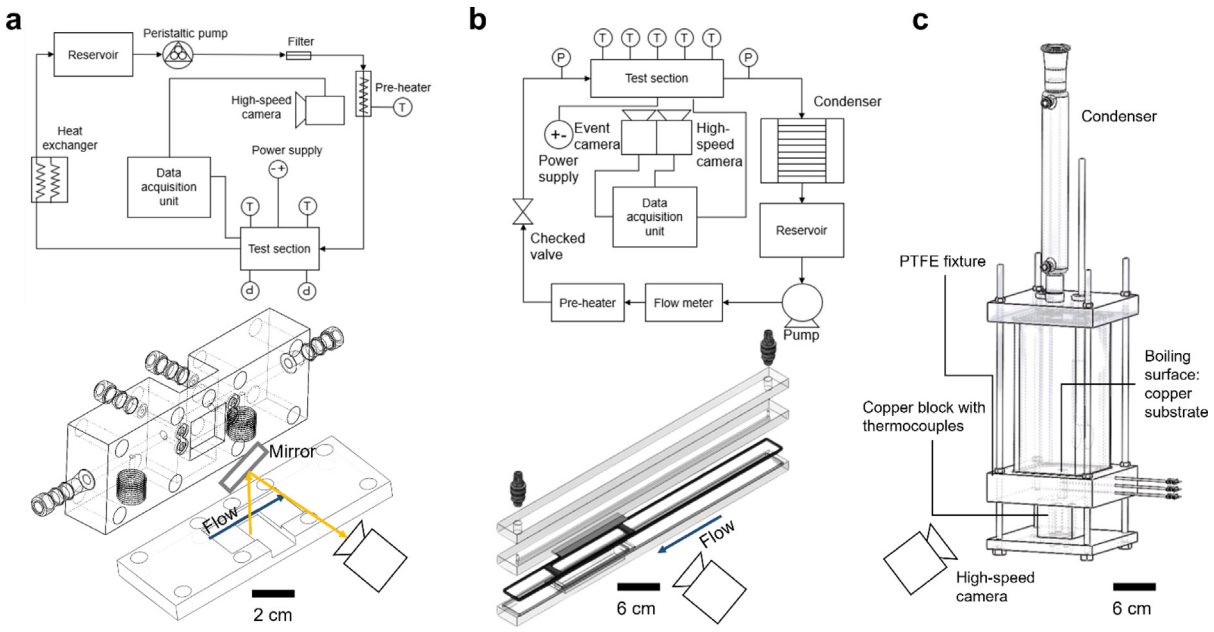

Fig. 2. Data Collection under Microflow Boiling, Flow Boiling, and Pool Boiling. (a) Microflow boiling setup: microchannels with a width of 400 μm and a depth of 100 μm heated up by a 12 mm × 12 mm heater. Lateral bubble movements are recorded through a cover glass from the top view. Heat flux is 2–3 W/cm^2, flow rate is 5 mL/min and inlet subcooling is around 20 °C. The working fluid is Flutec PP1. (b) Flow boiling setup: a polycarbonate channel with a rectangular cross-section and a heated length of 100 mm. Lateral and vertical bubble movements are observed through from the side view. Heat flux is 8–38 W/cm^2, flow rate is 430–640 mL/min, and inlet subcooling is around 4 °C. The working fluid is Flutec PP1. (c) Pool boiling setup: a copper surface (10 mm × 10 mm) soldered to a copper heating block. Vertical bubble dynamics are seen from the side view. Heat flux is 5–40 W/cm^2. The working fluid is DI water initially at room temperature. All videos are recorded using a high-speed camera (FAST-CAM Mini AX50) at 2,000 fps under atmospheric pressure (1 bar).

3.1. Event modality

Neuromorphic vision systems, such as event cameras, operate asynchronously and record only brightness changes instead of capturing complete frames. Each brightness change represents an event. Each event is encoded as pixel location, timestamp, and polarity (on or off), where each pixel is triggered independently. As a result, the data stream is highly sparse, which drastically reduces the amount of information that must be stored and transmitted. The efficiency allows each event to be timestamped independently at the microsecond scale, enabling temporal resolution and latency far beyond the limits of frame-based imaging. Critical transient phenomena like bubble nucleation, interface rupture, and coalescence in boiling can be monitored with minimal blur and delay. In addition, surface structures that obscures bubble boundaries and complicates segmentation can be eliminated. Such fine temporal resolution can be inherently provided by event cameras or be simulated through V2E algorithm [23] with the data from conventional high-speed video.

In this study, we apply V2E algorithm to generate the event modality (E) from previously recorded video. V2E detects inter-frame brightness variations above a defined threshold and converts them into timestamped events with pixel location and polarity. This procedure converts dense frame sequences into sparse, asynchronous event streams, thereby simulating the way neuromorphic vision systems operate without relying on dedicated hardware. The generated event-based modality preserves the key properties of events, data including high temporal precision, edge sensitivity, and sparse encoding of dynamic regions.

3.2. Reconstructed video modality

E2Vid converts events data back to conventional frame-based video (V) by using a recurrent neural network to interpolate and reconstruct intermediate frames from asynchronous event sequences [24]. This process restores intensity gradients and spatial details that are often lost in saturated or low-contrast high-speed recordings. By combining temporal sensitivity and spatial completeness, the reconstructed video capture dynamic features accurately. Since static backgrounds are not recorded by event sensors, the regenerated videos retain only dynamic bubble features, reducing data size without compromising essential information. This approach minimizes redundant content while complementing the sparse event encoding with spatially continuous details, producing dense intensity frames from simulated events.

3.3. Optical flow modality

In boiling under low flow rate or high heat flux conditions, numerous bubbles are continuously generated from the heating surface. These bubbles interact and overlap with one another, forming ambiguous boundaries that complicate the segmentation task. High-speed images, events representations from V2E, and reconstructed frames from E2Vid all ultimately collapse overlapping interfaces into a single composite contour, preventing reliable separation of individual bubbles. Optical flow is defined as the 2D vector field (e.g., direction and magnitude) describing apparent pixel-wise motion between consecutive frames based on variations in image brightness. It might be capable to address other modalities' limitations by inferring more detailed bubble dynamics features. By providing dense motion representations, it exposes subtle motion differences even within overlapped regions, enabling the network to disambiguate distinct bubbles and better capture their dynamics. Among existing optical flow models, GMFlow [25] leverages transformer-based global matching to establish long-range correspondences, enabling robust estimation of large displacements and reliable handling of occlusion. In this work, we apply GMFlow to convert conventional high-speed images into optical flow representations (O), where motion information – direction and magnitude – is visualized through color encoding.

4. Image fusion and instance segmentation

After multi-modality processing, we integrate E, V, O and H modality using two fusion strategies: pre-fusion and early-fusion depending

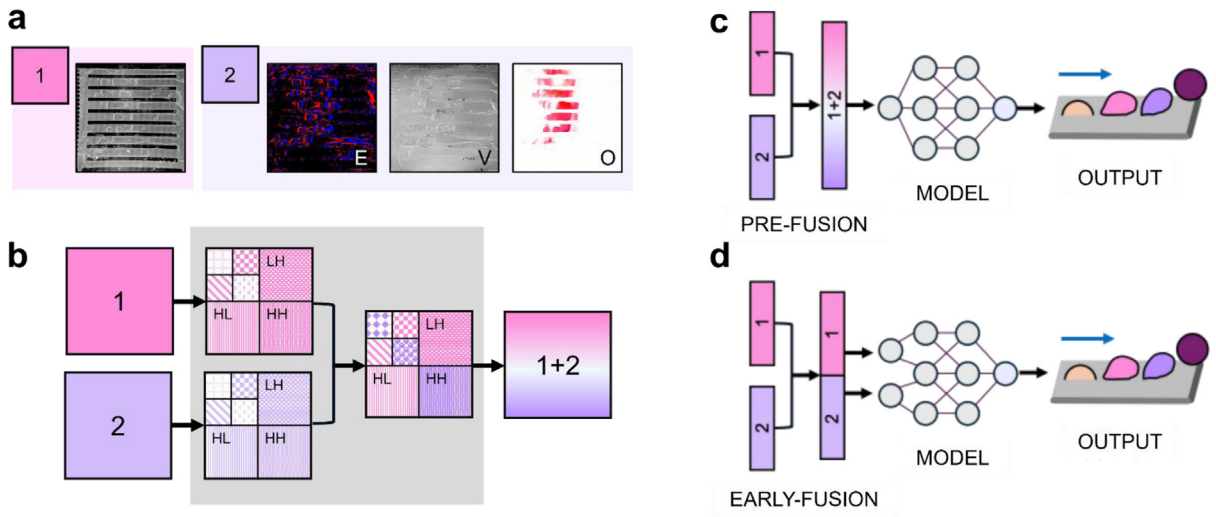

Fig. 3. Pre-fusion and Early-fusion Models for Multi-modal Instance Segmentation. (a) Multi-modality groups: group 1 contains H (high-speed images) while group 2 contains E (events), V (reconstructed video), and O (optical flow). Fusion is performed between modalities from different groups. (b) DWT process: two images from different modality group are transformed into the frequency domain, where corresponding sub-band coefficients are fused by weighted averaging (0.5/0.5). An inverse transform reconstructs a three-channel fused image that contains features from both modalities. (c) Pre-fusion model: the fused three-channel images obtained from (b) are directly used as network inputs, preserving the original YOLOv8 architecture without structural modification. (d) Early-fusion: two images from different modality group are concatenated to form a six-channel input, which is processed by a custom multi-modal module and a Squeeze-and-Excitation (SE) block that replace the first convolutional layer while other YOLOv8 structure remains unchanged.

on the stage at which fusion is applied. To provide supplementary information for high-speed images (H) while avoiding redundant or noisy features that could degrade segmentation accuracy, we focus on combinations that pair each auxiliary modality with H, as illustrated in Fig. 3a. Therefore, the integrations E+H, V+H, and O+H are selected

4.1. Pre-fusion

Pre-fusion combines modalities into a standard three-channel representation before being passed to the network, resulting in Pre-E+H, Pre-V+H, and Pre-O+H. This can be realized by applying discrete wavelet transform (DWT) [26–28], in Fig. 3b. DWT first transforms two images from the spatial domain to the frequency domain by decomposing them into sub-bands at each level. Fusion is then performed independently for each corresponding sub-band in each level across modalities by computing the weighted mean of their coefficients. Two-level decomposition is applied with equal weights of 0.5 assigned to each modality in order to ensure symmetric treatment and avoid introducing bias toward either source. After coefficient fusion, an inverse DWT (IDWT) transforms the fused coefficients back from the frequency domain to the spatial domain, resulting in one fused image. It allows multi-scale fusion and preserves important details from different modality [29]. In this process, information richness in input images is increased, which can potentially enhance the model's capacity to segment bubbles under small dataset. The resulting pre-fused images are then utilized as direct input for the YOLOv8 model [30], as shown in Fig. 3c. It should be noted that the YOLOv8 architecture follows a backbone-head paradigm with a Feature Pyramid Network (FPN) design [31] to extract multi-scale representations, while the head leverages a Path Aggregation Network (PANet) [32] to enhance feature fusion across different levels.

4.2. Early-fusion

Early-fusion combines modalities into a six-channel representation when being passed to the network, resulting in Early-E+H, Early-V+H, Early-O+H, and Early-H+H. In case of Early-H+H, the H modality is duplicated to form six channels and is used to compare fusion strategies. Unlike the pre-fusion method, early-fusion in Fig. 3d fuses two modalities directly within the segmentation model, YOLOv8. To do this, the first convolutional layer of YOLOv8 is replaced by a multi-modal module, which accept two modalities in the form of six channels input through separate branches. Each branch applies convolution to capture essential feature from its modality while reducing redundancy and noise. The outputs of the two branches are then concatenated into a single feature map and passed to a Squeeze-and-Excitation (SE) block [33]. The SE block evaluates the contribution of channel, adaptively assigns higher weights to informative one and lower weights to redundant ones. This adaptive weighting primarily serves to stabilize the six channels input, ensuring model focuses on the most relevant information. The reweighted features are subsequently propagated into the YOLOv8 backbone for instance segmentation.

4.3. Baseline model

Baseline models are constructed using the original YOLOv8 architecture, and each baseline is trained on the same dataset size as its corresponding fusion model. Three baselines (Fig. 4a–c) are defined for the individual experiments (microflow, flow, and pool boiling), with each trained on the corresponding high-speed images without fusion. In addition, one baseline is trained on a larger pool boiling dataset (Fig. 4(d) to serve as a reference for evaluating whether fusion can reduce the reliance on extensive annotations. In total, four baselines are considered in this study. Details of dataset size and sampling strategies are provided in Section 4.4.

4.4. Dataset size

This study consists of four training datasets: 100 images for three experimental setups (microflow boiling, flow boiling, and pool boiling) and 1,135 images for pool boiling. First, the 100 images are carefully selected to cover representative bubble behaviors, including nucleation, growth, coalescence, and departure, thereby reducing sampling bias and ensuring diverse instance dynamics within the limited annotation budget. All images are extracted from a single video at fixed heat flux and flow rate in each experiment (Fig. 4a–c). The training datasets are paired with their corresponding baselines (Section 4.3). In addition, for the pool boiling case, a larger dataset of 1,135 annotated images is

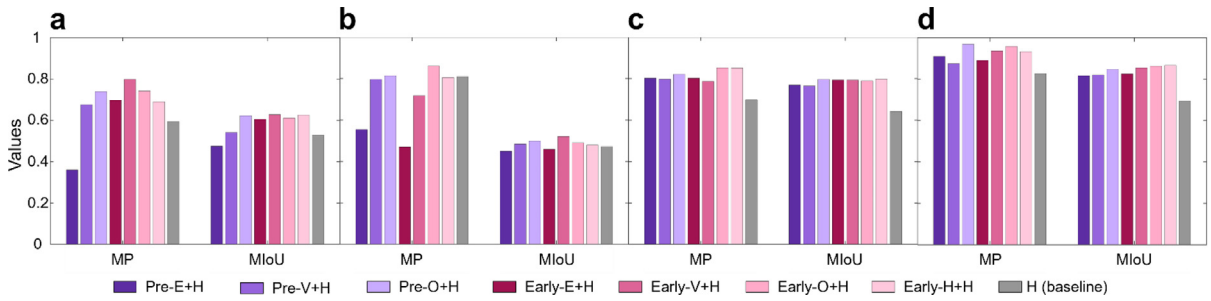

Fig. 4. Performance Results for Different Modality Combinations under Pre-fusion and Early-fusion Strategies. (a) Microflow boiling: Pre-O+H and Early-V+H show the highest segmentation accuracy, respectively, under pre- and early-fusion strategies. The best model (Early-V+H) outperforms the baseline by 34.4% in MP and 18.6% in MIoU. (b) Flow boiling (100 images): Pre-O+H and Early-O+H outperform other combinations in both pre- and early-fusion. Its best model (Early-O+H) achieves improvements of 6.3% in MP and 3.8% in MIoU compared with the baseline. (c) Pool boiling (100 images): similarly, Pre-O+H and Early-O+H show the highest accuracy. The best model (Early-O+H) shows improvements of 22.1% in MP and 22.9% in MIoU compared with the baseline. (d) Pool boiling (1,135 images): even with an expanded training dataset, Pre-O+H and Early-O+H remain the most effective in each fusion strategy. The best Early-O+H attains 95.8% in MP and 86.3% in MIoU, exceeding the baseline by 15.8% and 24.3%.

constructed from multiple videos under different heat flux conditions (Fig. 4(d). This larger dataset is used alongside its baseline to examine whether fusion strategies can compensate for reduced dataset size and approach the accuracy of models trained on much larger datasets. Annotation and training details are provided in *Supplementary Information Section* 3.

This study consists of three evaluation datasets: 50 sequentially annotated images for each boiling experiments. The annotations provide the ground truth, while the images serve as model inputs for prediction. These datasets are used exclusively for evaluation and are not involved in training, thereby ensuring an unbiased assessment of model performance across different experimental conditions.

5. Results

5.1. Evaluation

Evaluation Metrics. We have identified various evaluation metrics including mean precision (MP), mean recall (MR), mean F1 score (MF1), mean intersection over union (MIoU), and mean matching accuracy (MMA). These metrics are computed by comparing manually annotated ground truth with model predictions. For brevity, only two evaluation metrics, MP and MIoU, are shown in the main text. The other metrics and their definitions are provided in *Supplementary Information Section* 4 *and Table S2.*

The evaluation process begins by matching predicted bubble instances with ground truth instances using an IoU threshold (0.5) to define successful matches. This matching process yields instance-level true positives (TP), false positives (FP), and false negatives (FN), which quantify the model's segmentation performance [34]. For bubble segmentation, a true positive represents a predicted bubble mask that sufficiently matches a ground truth bubble, a false positive indicates a predicted bubble mask without a corresponding ground truth match, and a false negative denotes an undetected ground truth bubble.

Mean precision (MP): precision (P) measures the proportion of predicted bubble instances that correctly match ground truth bubbles, reflecting how many predicted detections are true positives.

$$P = TP/(TP + FP) \tag{1}$$

A high P indicates that most predicted bubbles correspond to actual bubbles, with few false detections. MP is calculated by averaging P across images in an evaluation dataset.

Mean IoU (MIoU): intersection over Union (IoU) measures the spatial overlap between a predicted bubble mask and the corresponding ground truth mask, defined as the ratio of their intersection to their union.

$$IoU = \text{intersection area}/\text{ union area} \tag{2}$$

A high IoU indicates that predicted bubbles closely align with the true bubble boundaries. MIoU is calculated by first averaging IoU over all bubbles within each image, and then averaging these image-level IoUs across an evaluation dataset.

Effectiveness of Fusion Method. Under the pre-fusion strategy, Pre-O+H demonstrates the best performance across all experiment setups. On average across the three experiments, it outperforms the other models by 28.2% in MP and 11.3% in MIoU. For early-fusion, Early-V+H performs best in microflow boiling, exceeding the other models by 12.6% in MP and 2.3% in MIoU, while Early-O+H achieves the best results in both flow boiling and pool boiling, outperforming the other models by an average of 20.8% in MP and 0.2% in MIoU.

A comparison of pre-fusion and early-fusion shows that the best-performing early-fusion model consistently surpasses its pre-fusion counterpart. Across the three experiments, early-fusion's best model delivers an improvement up to 8.0% in MP and 0.9% in MIoU over pre-fusion's best model. Relative to the baseline, the best early-fusion model shows increase up to 34.4% in MP and 18.6% in MIoU, whereas the best pre-fusion model achieves up to 24.5% and 17.6%.

Different Modalities. While some models with V show the best performance, the overall best performer is the model with the optical flow (O) modality. Discussed earlier, O shows the richest information that include bubble dynamics in direction and amplitude. It provides dense motion vectors that can disambiguate boundaries. It has the potential to provide the best observation regarding overlapping bubbles, which are common across different boiling experiments. This makes O a strong complementary source of information for H compared with other modalities, which supports accurate segmentation across complex bubble scenes. On the contrary, E shows relatively weaker performance. This can be attributed to the high sensitivity of the event modality. Since events are triggered by any brightness variation, the representation inevitably records changes unrelated to bubble dynamics, such as reflections or fluid disturbances. These non-bubble signals introduce noise and redundant features, which dilute the complementary value of E from the image fusion.

Impact of Channel Numbers. Early-H+H demonstrates 12.5% and 14.7% enhancements in MP and MIoU, respectively, when compared with baseline. Interestingly, Early-H+H also outperforms the baseline model although both models relay on only high-speed images only. While the baseline uses three-channel images, Early-H+H simply duplicates them to form a six-channel input. The better performance of Early-H+H implies that only increasing the input channel number can also bring better performance. The redundant input channels introduce no additional information but increases the dimensionality of the feature space, which enables the network's initial convolutional layers to learn more diverse filter combinations tailored to each channel. These

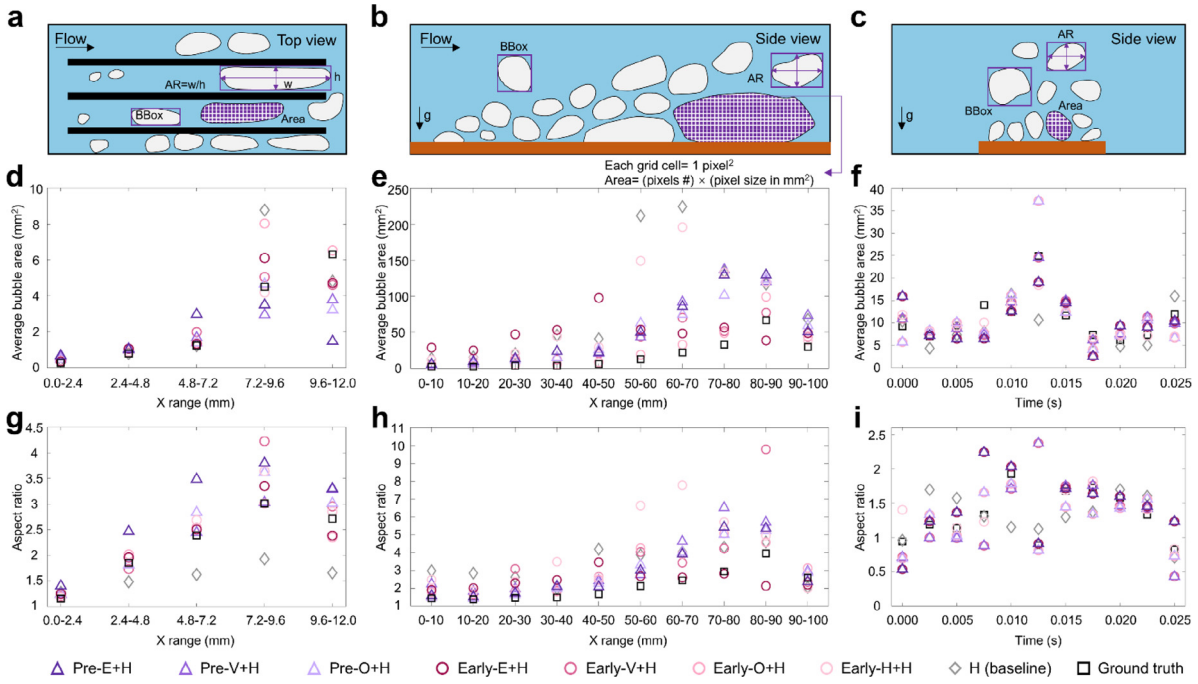

Fig. 5. Fusion Models with Ground Truth for Bubble Features. Bubble area is obtained by counting the pixels within the segmented mask, while the aspect ratio (AR) is defined as the width-to-height ratio of the bounding box (BBox). The first row illustrates how these quantities are measured for (a) microflow boiling, (b) flow boiling, and (c) pool boiling, respectively. Results for bubble area and aspect ratio are shown in (d–i). Bubble features are evaluated across spatial locations for microflow and flow boiling, and across time for pool boiling. Compared with the baseline, the best fusion model (Early-V+H for microflow boiling, Early-O+H for flow boiling and pool boiling) reduces prediction error in bubble area and bubble aspect ratio, respectively, by 18.1% and 5.9% for microflow boiling (d, g), by 92.6% and 157.8% for flow boiling (e, h), and by 107.1% and 82.3% for pool boiling (f, i).

learned combinations form richer low-level representations and facilitate the extraction of edges, textures, and object boundaries essential for instance-level discrimination.

Different Segmentation Models. Our previous study trained a Mask R-CNN model on 5,555 pool boiling images [35], achieving an average value of 89.0% in terms of precision. When the training set for the Mask R-CNN is reduced to 1,135 images, performance decreased compared to the model trained on 5,555 images, with MP at 82.7% (baseline in Fig. 4(d). However, as illustrated in Fig. 4d, the YOLOv8 model trained on the smaller 1,135 image dataset with multi-modal inputs reaches 95.8% in MP, surpassing the single modality model trained on the larger dataset. This result highlights the potential of proposed approaches to reduce the need for large datasets and extensive manual labeling, while still maintaining or even improving segmentation accuracy.

5.2. Physics-rich features

The trained models are then applied to extract quantitative and physically meaningful bubble features, including bubble area and aspect ratio, for different boiling conditions. In microflow boiling, bubbles are confined and slide along the heated wall. Although they detach from their original locations, the vapor structures continue to influence surface heat transfer. In this regime, the void fraction describes the vapor–liquid portion that affects the local heat transfer coefficient, and the aspect ratio reflects deformation caused by geometric confinement and bubble interactions. In flow boiling and pool boiling, the measured bubble area provides equivalent diameters and growth curves used in RPI heat transfer analyses. The aspect ratio indicates flow regime transitions that shape heat transfer behavior in flow boiling and also serves as a useful morphological CHF indicator in pool boiling.

As illustrated in Fig. 5a–c, the bubble area is determined by the pixel count of the segmented mask, and the aspect ratio is defined as the ratio of bubble width to height of the bounding box (bbox). The bubble features extracted by various models are evaluated across spatial locations for microflow and flow boiling, and across time for pool boiling with ground truth in Fig. 5 d–i.

The best fusion models – Early-V+H for microflow boiling, and Early-O+H for flow and pool boiling – exhibit an average difference between prediction and ground truth of 30.5% ±16.1% for bubble area and 14.0% ± 29.9% for bubble aspect ratio across the three experiments. In contrast, when other models are used, the average differences increase to 141.5% ± 212.6% and −29.5% ± 14.3%, respectively. Compared with the baseline, the best fusion model reduces prediction error in bubble area and bubble aspect ratio, respectively, by up to 92.6% and 157.8%. These results demonstrate that selecting an appropriate segmentation model enables accurate bubble data extraction even with limited training data. Our boiling experiments cover a wide range of operating conditions, producing bubbles of varying sizes and shapes, which makes the task particularly challenging. Despite this diversity, the proposed multi-modal segmentation framework consistently achieves high segmentation accuracy, highlighting its robustness across different experimental conditions.

6. Conclusion

The main motivation behind this paper is to solve the need of large dataset under the requirement for high segmentation accuracy with minimum training data. In this study, we propose a framework that leverages fusion of multiple modalities to maintain segmentation accuracy even with a limited number of datasets. Modalities, such as events, reconstruction video from events, and optical flows, are generated from high-speed images through an image processing pipeline. The use of multiple modalities benefits the segmentation model by providing additional information regarding bubble trajectory and velocity. For this, we develop two fusion strategies: pre-fusion and early-fusion depending on whether the two modality is fused before

or at the stage of the segmentation. The best fusion models consistently (Early-V+H for microflow boiling, Early-O+H for flow boiling and pool boiling) achieve improvements up to 20%–40% compared to the baseline model. Except for numerical improvements, the proposed framework enables more accurate quantification of bubble dynamics in phase-change heat transfer, reducing prediction errors to about 20% for bubble area and aspect ratio relative to the ground truth.

Beyond techniques adopted in our pipeline, modality may be obtained from motion history images and spatiotemporal filtering, and includes more feature diversity. On the fusion side, while our early-fusion strategy focuses on front-layer concatenation, future work may investigate staged integration strategies. For example, middle fusion selectively combines modality-specific features after a few backbone layers, while late fusion merges task-specific features just before prediction. Furthermore, channel-level weighting in SE block might suppress an entire modality if its signals are weak but still important. Adaptive modality weighting could be introduced to balance the contribution of each modality during training. One possible realization is to aggregate the three channels from each modality using global average pooling, and then the modality-specific weights is generated by a softmax-based gating module.

These directions would enable more flexible fusion architectures, potentially leading to novel multi-modal segmentation networks applied across different boiling regimes and diverse domains. The consecutive instance-level masks provide the foundation for tracking, which enables dynamic parameters extraction. Building on this, a tracking framework will be developed to construct nucleation, growth and departure maps that characterize bubble dynamics across the heating surface. In addition, the framework's ability to preserve accuracy across boiling regimes further highlights its potential applicability to other two-phase flow systems, such as condensation, immersion cooling, and fuel-cell liquid-gas interfaces. Moreover, the multi-modality framework can be extended beyond thermal-fluid sciences to fields such as biological imaging and remote sensing, where multi-scale, high-speed, and high-fidelity segmentation is equally critical.

Table of abbreviations

Abbreviations	*Definition*
AR	aspect ratio
BBox	bounding box
CHF	critical heat flux
CT	computed tomography
DWT	discrete wavelet transform
E2Vid	events-to-video reconstruction
FN	false negatives
FP	false positives
FPN	Feature Pyramid Network
GMFlow	global matching flow
IoU	intersection over union
MA	matching accuracy
Mask R-CNN	mask region-based convolutional neural network
MF1	mean F1 score
MIoU	mean intersection over union
ML	machine learning
MMA	mean matching accuracy
MP	mean precision
MR	mean recall
MRI	magnetic resonance imaging
ONB	onset of nucleate boiling
P	precision
PANet	Path Aggregation Network
PET	positron emission tomography
R	recall
RPI	Rensselaer Polytechnic Institute
SE	Squeeze-and-Excitation block
TP	true positives
V2E	video-to-events
YOLO	you only look once

CRediT authorship contribution statement

Xiaojing Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sanghyeon Chang:** Writing – review & editing, Data curation. **Youngjoon Suh:** Writing – review & editing, Data curation. **Yoonjin Won:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no competing financial interests that influenced the work reported in this paper.

Acknowledgments

X.Z. expresses gratitude for the graduate fellowship received from the UCI Department of Mechanical and Aerospace Engineering. X.Z., S.C., Y.S., and Y.W. also gratefully acknowledge funding support from the Office of Naval Research (ONR) under Grant No. N00014-22-1-2063, and MURI Grant No. N00014-24-1-2575 with Dr. Mark Spector serving as the program officer.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.aitf.2026.100028.

Data availability

Data will be made available on request.

References

[1] N. Basu, G.R. Warrier, V.K. Dhir, Onset of nucleate boiling and active nucleation site density during subcooled flow boiling, J. Heat Transf. 124 (4) (2002) http://dx.doi.org/10.1115/1.1471522, Art. (4).

[2] Y. Liu, Y. Du, G. Fei, T. Zhou, C. Zhao, Effects of surface wettability on bubble departure and critical heat flux: A parametric study based on 3-D dynamic force analysis model, Int. J. Therm. Sci. 186 (2023) 108149, http://dx.doi.org/10.1016/j.ijthermalsci.2023.108149.

[3] C.-N. Huang, S. Chang, Y. Suh, I. Mudawar, Y. Won, C.R. Kharangate, Machine learning boiling prediction: From autonomous vision of flow visualization data to performance parameter theoretical modeling, Int. J. Multiph. Flow 179 (2024) 104928, http://dx.doi.org/10.1016/j.ijmultiphaseflow.2024.104928.

[4] K. Necdet, M.Z. Podowski, On the modeling of multidimensional effects in boiling channels, in: Presented At the National Heat Transfer Conference, Minneapolis, Minnesota, USA, 1991.

[5] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask R-CNN, in: 2017 IEEE International Conference on Computer Vision, ICCV, IEEE, Venice, 2017, pp. 2980–2988, http://dx.doi.org/10.1109/ICCV.2017.322.

[6] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, 2015, http://dx.doi.org/10.48550/ARXIV.1505.04597, arXiv.

[7] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8, 2023, Ultralytics. [Online]. Available: https://github.com/ultralytics/ultralytics.

[8] Y. Kim, H. Park, Deep learning-based automated and universal bubble detection and mask extraction in complex two-phase flows, Sci. Rep. 11 (1) (2021) 8940, http://dx.doi.org/10.1038/s41598-021-88334-0.

[9] J.H. Seong, M. Ravichandran, G. Su, B. Phillips, M. Bucci, Automated bubble analysis of high-speed subcooled flow boiling images using U-net transfer learning and global optical flow, Int. J. Multiph. Flow 159 (2023) 104336, http://dx.doi.org/10.1016/j.ijmultiphaseflow.2022.104336.

[10] R. Sapkota, D. Ahmed, M. Karkee, Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments, Artif. Intell. Agric. 13 (2024) 84–99, http://dx.doi.org/10.1016/j.aiia.2024.07.001.

[11] T.T. Thai, et al., Comparative analysis of stomatal pore instance segmentation: Mask R-CNN vs. YOLOv8 on phenomics stomatal dataset, Front. Plant Sci. 15 (2024) 1414849, http://dx.doi.org/10.3389/fpls.2024.1414849.
[12] S. Chang, et al., BubbleMask: Autonomous visualization of digital flow bubbles for predicting critical heat flux, int, J. Heat Mass Transf. 217 (2023) 124656, http://dx.doi.org/10.1016/j.ijheatmasstransfer.2023.124656.
[13] Z. Chen, H. Hou, Improved YOLOv8 algorithm for bubble detection in stainless steel welded pipe airtightness testing, in: X. Zhang, M. Yin (Eds.), Fifth International Conference on Computer Vision and Data Mining, ICCVDM 2024, SPIE, Changchun, China, 2024, p. 62, http://dx.doi.org/10.1117/12.3048228.
[14] Z. Yang, et al., A deep learning-based segmentation method for multi-scale and overlapped bubbles in gas–liquid bubbly flow, Phys. Fluids 37 (3) (2025) 033343, http://dx.doi.org/10.1063/5.0254679.
[15] Y. Cao, D. Pang, Y. Yan, Y. Jiang, C. Tian, A photovoltaic surface defect detection method for building based on deep learning, J. Build. Eng. 70 (2023) 106375, http://dx.doi.org/10.1016/j.jobe.2023.106375.
[16] J. Huang, K. Zeng, Z. Zhang, W. Zhong, Solar panel defect detection design based on YOLO v5 algorithm, Heliyon 9 (8) (2023) e18826, http://dx.doi.org/10.1016/j.heliyon.2023.e18826.
[17] D. Wan, et al., YOLO-MIF: Improved YOLOv8 with Multi-Information fusion for object detection in Gray-scale images, Adv. Eng. Inf. 62 (2024) 102709, http://dx.doi.org/10.1016/j.aei.2024.102709.
[18] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, L. Huang, What makes multi-modal learning better than single (provably), 2021, http://dx.doi.org/10.48550/ARXIV.2106.04538, arXiv.
[19] T. Zhou, S. Ruan, S. Canu, A review: Deep learning for medical image segmentation using multi-modality fusion, Array 3–4 (2019) 100004, http://dx.doi.org/10.1016/j.array.2019.100004.
[20] J. Wagner, V. Fische, M. Herman, S. Behnk (Eds.), Multispectral pedestrian detection using deep fusion convolutional neural networks, presented at the ESANN, louvain-la-neuve, in: Belgique: 24th European Symposium on Artificial Neural Networks, in: Computational Intelligence and Machine Learning, 2016.
[21] Y.-L. Hou, Y. Song, X. Hao, Y. Shen, M. Qian, Multispectral pedestrian detection based on deep convolutional neural networks, in: 2017 IEEE International Conference on Signal Processing, Communications and Computing, ICSPCC, IEEE, Xiamen, 2017, pp. 1–4, http://dx.doi.org/10.1109/ICSPCC.2017.8242507.
[22] J. Liu, S. Zhang, S. Wang, D.N. Metaxas, Multispectral deep neural networks for pedestrian detection, 2016, http://dx.doi.org/10.48550/ARXIV.1611.02644, arXiv.
[23] Y. Hu, S.-C. Liu, T. Delbruck, V2e: From video frames to realistic DVS events, 2020, http://dx.doi.org/10.48550/ARXIV.2006.07722, arXiv.
[24] H. Rebecq, R. Ranftl, V. Koltun, D. Scaramuzza, High speed and high dynamic range video with an event camera, 2019, http://dx.doi.org/10.48550/ARXIV.1906.07165, arXiv.
[25] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, D. Tao, GMFlow: Learning optical flow via global matching, 2021, http://dx.doi.org/10.48550/ARXIV.2111.13680, arXiv.
[26] D.E. Krebs, J.E. Edelstein, S. Fishman, Reliability of observational kinematic gait analysis, Phys. Ther. 65 (7) (1985) 1027–1033, http://dx.doi.org/10.1093/ptj/65.7.1027.
[27] A. Cohen, I. Daubechies, J.-C. Feauveau, Biorthogonal bases of compactly supported wavelets, 45, (5) 1992, pp. 485–560, http://dx.doi.org/10.1002/cpa.3160450502.
[28] A. Cohen, I. Daubechies, B. Jawerth, P. Vial, Multiresolution analysis, wavelets and fast algorithms on an interval, Comptes Rendus Académie Sci. Sér. 1 Mathématique 316 (5) (1993) 417–421.
[29] M. Li, W. Chen, T. Zhang, Classification of epilepsy EEG signals using DWT-based envelope analysis and neural network ensemble, Biomed. Signal Process. Control. 31 (2017) 357–365, http://dx.doi.org/10.1016/j.bspc.2016.09.008.
[30] YOLO by Ultralytics. [Online]. Available: https://github.com/ultralytics/ultralytics.
[31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, 2016, http://dx.doi.org/10.48550/ARXIV.1612.03144, arXiv.
[32] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, 2018, http://dx.doi.org/10.48550/ARXIV.1803.01534, arXiv.
[33] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, 2017, http://dx.doi.org/10.48550/ARXIV.1709.01507, arXiv.
[34] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. 27 (8) (2006) 861–874, http://dx.doi.org/10.1016/j.patrec.2005.10.010.
[35] Y. Suh, et al., VISION-iT: A framework for digitizing bubbles and droplets, Energy AI 15 (2024) 100309, http://dx.doi.org/10.1016/j.egyai.2023.100309.