Classification of time-variant transcritical droplets with attention-enhanced deep spatiotemporal learning

Alice Evelyn Ashby*[1], Khuong An Nguyen[2], Andy Philippides[3], Julien Manin[4], Cyril Crua[1]

[1] Energy and Materials Engineering Research Centre, University of Sussex, Brighton, UK
[2] Department of Computer Science, Royal Holloway, University of London, Surrey, UK
[3] Sussex AI, University of Sussex, Brighton, UK
[4] Combustion Research Facility, Sandia National Laboratories, Livermore, USA
*Corresponding author: A.Ashby@sussex.ac.uk

Abstract

The transcritical mixing of liquid fuel sprays is a process characterised by a fuel droplet's transition from a classical evaporation state to a state of diffusive mixing. Although we previously proposed a phenomenological model identifying distinct mixing regimes (classical evaporation, transitional mixing, and diffusive mixing), analysing such large image datasets still requires significant human intervention. This is primarily due to these mixing regimes being identified through temporal criteria (i.e. evolution of droplet shape in time), which is particularly challenging to automate using traditional image processing algorithms. To address this issue, we designed a deep spatiotemporal learning algorithm trained on human-annotated frames of synthetic transcritical droplets, inspired by high-speed long-distance microscopy videos. In this paper we present our two-stage detection and classification pipeline, where a multiple-object tracking (MOT) algorithm based on YOLOv11 and an integrated BoT-SORT tracking layer initially detect moving droplets and isolate them at the object level. Then, a novel residual convolutional neural network and bidirectional long short-term memory network with a temporal attention module (CNN-BiLSTM-TAM) is proposed to classify the mixing regimes using the extracted object-level droplet images. Our algorithm is designed to learn both the rich visual characteristics of the droplets and their time-based evolution, using spatial and temporal attention to capture the most informative frames in the droplet image sequences. We provide robust empirical validation of our work through attention map visualisation and performance comparison with two state-of-the-art image classifiers, showcasing improvements in precision, recall, and F1 score of +50%, +31%, +40% and +81%, +56%, +72%.

Keywords: transcritical mixing, multiple object tracking, video classification, spatiotemporal learning

1 Introduction

Over the last several years, there has been a substantial effort to investigate evaporation and mixing regimes for fuels at conditions relevant to combustion-ignition engines and aero-engine combustors, particularly with the interests of ensuring efficient combustion and reduced toxic emissions, motivated by stringent new regulations to be implemented globally. A focal point of such research is that of the transcritical mixing of liquid fuel sprays, a process characterised by a fuel droplet's transition from a classical evaporation state to a diffusive mixing state [1]. Previously, we proposed a phenomenological model identifying three distinct mixing regimes from high-speed long-distance microscopy videos:

- **Classical evaporation** Droplets are characterised by significant surface tension and maintain a spherical shape. A d^2 law-based evaporation model can be assumed for their lifetime [1].
- **Transitional mixing** Droplets initially maintain surface tension, then exhibit rapidly accelerated vaporisation and stretching starting at the wake side of the droplet. They transition from a spherical shape to a more stretched and deformed profile, continuously oscillating until eventual vaporisation. Towards the end, they may resemble a backwards-facing bag [1].
- **Diffusive mixing** Droplets may initially be characterised by surface tension but quickly experience a complete (or near complete) loss of this, where two-phase flow cannot be assumed any more. The droplet will quickly lose its spherical shape, oscillating and deforming continuously [1].

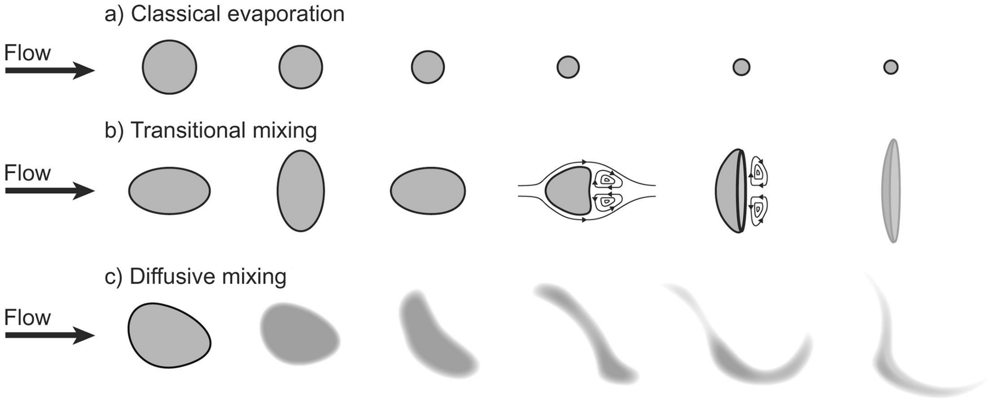


Figure 1. *A conceptual model of droplet morphological evolution based on mixing regimes [1].*

At present, identifying such regimes relies on building thermodynamic maps to define the operating conditions for which the dominant regime shall be considered classical, transitional, or diffusive, supported by human classification based on temporal criteria (see Figure 1) [1]. As such, accurate analysis of optical data at scale is an arduous task, often requiring significant human intervention. In recent years, researchers across our teams have attempted to automate the process through machine learning (ML) modelling, but this has proven to be a challenging endeavour in part due to the large volume of data, lack of standardisation, and the innate temporal nature of the droplet's themselves.

We provide the first preliminary investigation into using ML models for analysis of high-speed long-distance microscopy videos of transcritical mixing under operating environments typical of cruise and take-off conditions [2]. We employ the single-pass object detector YOLOv5 [3] on a human-annotated dataset of 577 droplets from the experiments in [1, 2], and found that for cruise conditions with an ambient pressure of 2 MPa, there was reasonable agreement observed between YOLOv5 and human classification. However, for take-off conditions at 6 MPa, there was less agreement, seen to be fuel-dependant (for example, Jet-A had less agreement than bicyclohexyl) [2]. We note two limitations of our analysis; we omit the transitional due to lack of observations, and some classical droplets may have been misclassified as diffusive due to the prevalence of unfocused droplets and potential anisotropy in the illumination due to beam steering [2]. Later, we follow-up with multi-class classification using the multi-pass object segmentor Mask R-CNN [4] on a dataset of 1,515 droplets from the dataset in [2]. Similarly to [2], at 2 MPa reasonable agreement was observed, but this degraded at 6 MPa, with many transitional and diffusive droplets misclassified as the classical regime [5]. These results provided a promising attempt at accelerating the analysis of large transcritical video datasets with deep learning.

This manuscript presents a continuation of our previous work with an approach from a different perspective. In the previous works, static models were proposed to classify the mixing regimes in the spatial dimension, and the temporal dimension was not considered. Spatiotemporal patterns are spatial patterns that evolve along the temporal dimension, and often contain rich contextual information that can inform classification, such as object shape and motion. A sequence of images, such as frames in a video, often encode spatiotemporal patterns. Static networks that are unaware of the temporal dimension thus make the incorrect implicit assumption that the frames are independent and identically distributed, therefore, the current frame must be independent of the frame before it. This is often not the case with frames being highly correlated in time. As such, adjacent frames should instead be considered in conjunction, and not doing so can lend to unreliable classifications. As such, we propose a two-stage detection and classification pipeline, capable of learning the full spatiotemporal patterns embedded in the video data.

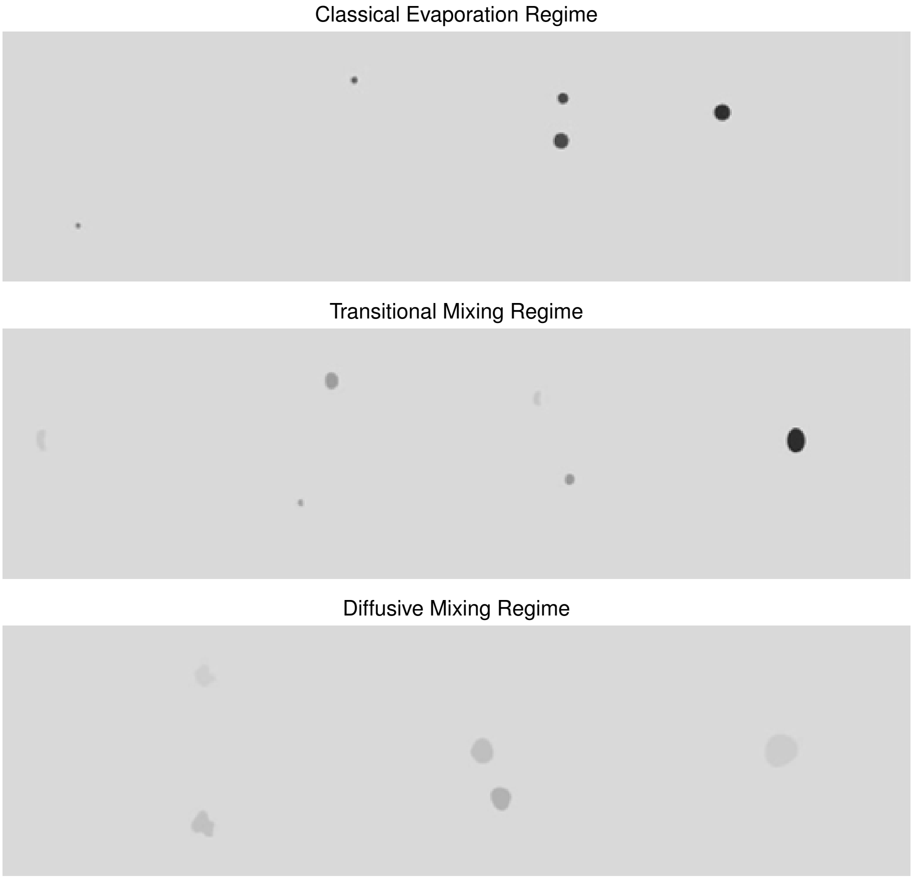
The rest of the paper is organised as follows: Section 2 provides a brief overview of the challenges observed in the data through our previous studies with YOLOv5 and Mask R-CNN, and the synthetic dataset created to control such challenges in our experiments. Section 3 outlines the first stage of our pipeline and results, in which the YOLOv11 object detector with the integrated BoT-SORT tracker is employed to create an object-level dataset. Section 4 outlines the subsequent second stage of our pipeline and results, where the object-level dataset is used to train and validate a spatiotemporal network CNN-BiLSTM-TAM. Lastly, Section 5 concludes the key results of our paper in a concise manner.

2 Synthetic Dataset

As outlined in Section 1, there are numerous challenges associated with accurate detection and classification of transcritical droplets in liquid fuel sprays for vision models, and we hypothesise that such issues will affect our proposed tracking step as well. These challenges are summarised below:

- **Unfocused droplets** Unfocused droplets may be wrongly identified as evaporating or diffusing droplets due to visual similarities, such as a loss in gradient and a less defined boundary.
- **Motion blur** Fast moving droplets captured immediately exiting from the injector may exhibit motion blur even when in focus by the optical equipment.
- **Occlusion** Sprays typically have many hundreds of droplets per frame, which can occlude each other in the crowded environment, especially at elevated ambient pressures such as 6 MPa.
- **Noise artefacts** Optical quality can be reduced by grainy textures due to vapour and other noise artefacts during injection, which can also cause misclassification with diffusive droplets, etc.
- **Small droplet diameter** Many droplets are <100 µm and are difficult to detect by vision models.
- **Gradient blur** A blurring of the gradient can be seen with evaporating or diffusing droplets, making detection more complex, especially with the aforementioned noise artefacts.

As addressing each of these issues in turn is a complex and arduous task, we opted to focus on controlling the prevalence and intensity of these challenges in our training set. To achieve this, we designed a transcritical droplet simulation in MATLAB to generate spray videos (shown in Figure 2) inspired by high-speed long-distance microscopy datasets, particularly the SAF22 dataset used in both [2, 5].



Classical Evaporation Regime

Transitional Mixing Regime

Diffusive Mixing Regime

Figure 2. *Synthetic spray images of the three transcritical mixing regimes. The injector region on the right and image height is cropped for clarity.*

We focus on modelling motion blur, occlusion, small droplet diameters, and gradient blur in our synthetic videos. However, we were unable to adequately model unfocused droplets and noise artefacts, due to the inherent nature of those challenges being associated with the optical quality and experimental setup. Our synthetic dataset is split 80/20 into train and validation sets respectively, with the train set consisting of 1,976 images with 5,221 droplet instances (by class: 2,184 classical, 1,650 transitional, 1,387 diffusive) and the validation set 493 images with 1,337 droplets (by class: 550 classical, 453 transitional, 334 diffusive). Background images (with no droplets) make up approx. 3.9% of the dataset. All droplet instances were human-annotated using the Computer Vision Annotation Tool (CVAT). We plan to make this droplet simulation and annotated dataset publicly available in due time.

3 Stage 1 - Detection and Tracking Network

3.1 Methodology

The first stage of our detection and classification pipeline uses the state-of-the-art object detector YOLOv11 [6] with improved detection for small and occluded objects, attributed to new spatial attention blocks to support the existing multi-scale prediction architecture, and reductions in training computational overhead [7]. We opt for a medium scale architecture with 253 layers and 22,420,896 learnable parameters, pretrained on the Microsoft COCO dataset. The core model hyperparameters initial and final learning rate, L2 regularisation, and epoch warmup, are tuned using tree-structured parzen estimators (TPE) [8] for 30 iterations to determine which set of hyperparameters exhibit the highest detection performance. To speed up tuning, we use automated early-stopping to prune unpromising iterations using the asynchronous successive halving algorithm (ASHA) [9] in a simplified and balanced configuration.

The training dataset is augmented using hue, saturation, brightness, rotation, translation, scale, horizontal flip, and mosaic augmentation. In practise, high-speed long-distance microscopy images are not standardised, often captured with different optical setups, which can affect the image resolution, lighting, jet flow direction, camera focus, and other properties. Data augmentation can aid the model in generalising despite these variations, through simulating different lighting conditions and exposure levels, droplet orientations and scales, and jet flow direction. We specifically do not use perspective or shear augmentation, in order to retain essential morphological features of the droplets, such as sphericity. The data augmentation parameters were tuned with the same methodology as the model hyperparameters.

The YOLOv11 model with the best hyperparameters obtained through the aforementioned tuning process is integrated with the tracker BoT-SORT [10]. YOLOv11 detects droplets based on spatial features derived from visual characteristics, and then BoT-SORT persistently tracks these detected droplets across sequential frames of the simulated spray videos using a unique ID assigned to each droplet. BoT-SORT aids the model in linking detections across frames to a single droplet, allowing us to crop the contents of each bounding box of which the droplet is contained and export into an object-level image dataset. Tracking by detection using an ID for each droplet enables us to maintain sequential order and therefore preserve the temporal consistency across frames. After cropping, bounding boxes are resized to 224x224, as this is the optimal size for ResNet models pretrained on the ImageNet dataset (see Section 4). To preserve aspect ratio, we utilise padding with average pixel intensity. We then apply Gaussian blur to smooth aliasing artefacts for better temporal consistency. Prior to exporting, short-lived lower-confidence tracks are filtered out, ensuring only strong and stable tracks are retained.

3.2 Results

Hyperparameter optimisation took 20 hours, and model training with the best hyperparameters and cosine annealing took 1 hour for a maximum of 200 epochs, on a RTX 3090 Ti GPU. Described in Section 2, training images have a resolution of 894x486, but YOLOv11 prefers values divisible by 32. Therefore, we resize them to 896x512, which is closest to the native resolution. Due to the large image resolution, a batch size of 16 was selected to work with the 24 GB VRAM. The optimiser AdamW [11] is utilised due to its improved performance and faster convergence for medium-to-large-scale models. Early-stopping after 10 epochs is used to prevent overfitting on the training set.

Table 1 summarises the results of our model on the validation set, compared to our YOLOv5 [2]. Only detections with a confidence value of $\geq$50% are considered in [2], as such, we show our results at $\geq$50% confidence. Table 2 summarises the class-wise results on the validation set, compared to our previous work. Figure 3 depicts the F1-confidence curve and the precision-recall (PR) curve of our YOLOv11.

In Table 1, we observe that YOLOv11 achieves better performance across all metrics. A precision of 1 indicates that there were no false positives, and all predictions made were correct, whereas for YOLOv5, 38% of predictions were incorrect. A recall of 0.81 indicates that 81% of actual ground truth droplets were detected; therefore, there were only 19% of the droplets missed, compared to 55% by YOLOv5. Lastly, mAP@50-95 provides a comprehensive view of model performance across different levels of detection difficulty; for YOLOv11, mAP@50-95 increases by 57% over YOLOv5 indicating far better detection capability across all three classes. In Table 2, we observe YOLOv11 achieves much

Table 1. *Detection results of YOLOv11 on the validation dataset compared to YOLOv5. mAP is mean average precision at various intersection over union (IoU) thresholds.*

	Precision	Recall	mAP@50	mAP@75	mAP@50-95
YOLOv5 [2]	0.62	0.45	0.51	-	0.22
YOLOv11 (ours)	**1.00**	**0.81**	**0.91**	**0.89**	**0.79**

Table 2. *Class-wise detection results of YOLOv11 on the validation dataset compared to YOLOv5 and Mask R-CNN. AP is average precision.*

	Classical Class AP	Transitional Class AP	Diffusive Class AP
YOLOv5 [2]	0.46	0.61	0.48
Mask R-CNN [5]	0.56	0.49	0.47
YOLOv11 (ours)	**0.64**	**0.85**	**0.88**

better performance for the transitional and diffusive regimes over YOLOv5 and Mask R-CNN, and for the classical regime over YOLOv5. However, YOLOv11 only marginally improves the performance for the classical regime over Mask R-CNN, as this is the class which YOLOv11 struggled with the most.

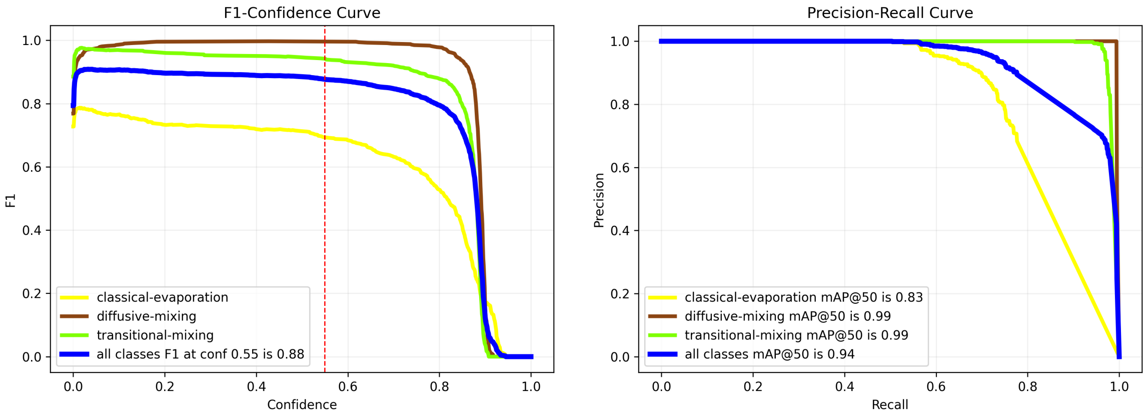


Figure 3. *F1-confidence and precision-recall (PR) curves describing the detection performance of YOLOv11. The F1 score is the harmonic mean of precision and recall.*

The F1-confidence curve in Figure 3 describes the relationship between the F1 score and model confidence threshold for each class. It indicates that the diffusive mixing regime is highly stable, maintaining a high F1 insensitive to thresholds, while transitional mixing performs well with an average F1 of 0.95, though with a steeper decline as the threshold increases. In contrast, classical evaporation has a lower peak F1 around 0.78 with rapid deterioration in performance past a threshold of 0.5. At a global threshold of 0.55 (marked with the dotted red line), a mean F1 of 0.88 is observed. As such, we use 0.55 as our minimum confidence threshold when generating the object-level dataset, as we believe it to be an acceptable performance trade-off. The PR curve in Figure 3 illustrates the trade-off between precision and recall across different model confidence thresholds. It shows strong and robust performance for both diffusive and transitional mixing, each achieving class-wise mAP@50 of 0.99, with their curves remaining near-perfect until very high recall levels, indicating that the model makes very few false positives or false negatives for these classes. In contrast, classical evaporation achieves a lower mAP@50 of 0.83, and its PR curve begins declining past 0.7 recall, indicating increased model uncertainty.

The violin plot in Figure 4 illustrates the distribution of detection confidence scores across the three transcritical mixing regimes. Classical evaporation exhibits a wide, multimodal distribution, with a peak near 0.9 and a substantial tail extending toward lower confidence scores. This suggests a high variance in prediction certainty. In contrast, diffusive mixing has a narrow and sharply peaked distribution, centred around 0.88-0.9 with minimal spread, indicating high and consistent confidence. Lastly, transitional mixing shows a slightly skewed distribution, similar in shape to that of classical evaporation but shifted upward towards higher confidence. Whilst it shows a broader confidence range than diffusive mixing, it

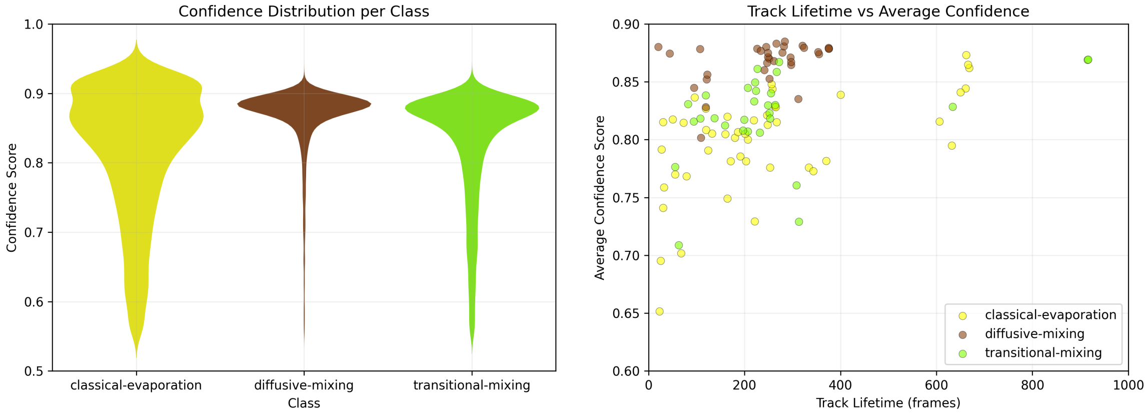


Figure 4. *Class-wise confidence distribution and track lifetime of YOLOv11 and BoT-SORT.*

still peaks around 0.88, suggesting a moderate consistency with occasional uncertainty. The scatter plot in Figure 4 illustrates the relationship between frame-wise track persistence and the average confidence score per track for each class. Here, we can see that diffusive mixing tracks exhibit the highest overall confidence and consistency, clustered tightly around confidence scores of 0.86-0.89. Classical evaporation displays a wider confidence spread, with many tracks having an average confidence near 0.8 for both short and long-lived tracks. Transitional mixing tracks tend to cluster slightly higher in confidence than classical evaporation tracks, with some long-lived tracks reaching confidence levels above 0.85.

The scatter plots on the left in Figure 5 explore the relationship between bounding box area and model confidence for each class. For classical evaporation, the model exhibits strong size sensitivity, with high-confidence predictions concentrated among larger bounding boxes. This highlights a potential weakness in detecting small-scale classical droplets, likely due to reduced resolution. For diffusive mixing, the model demonstrates robust performance regardless of droplet scale, suggesting that the visual characteristics of this regime are highly learnable. For transitional mixing, the model shows increased uncertainty across small to mid-sized bounding boxes, with broader distribution of confidence suggesting transitional droplets may be inherently harder to characterise, potentially due to visual inter-class similarity or ambiguity. The scatter plots on the right in Figure 5 explore the relationship between visual clarity (using the standard deviation of the greyscale cropped bounding box as a contrast score) and model confidence for each class. For classical evaporation, model confidence is positively influenced by visual clarity, indicating that clearer, higher-contrast droplets enable more reliable detection, whereas with lower-contrast droplets, detection is less reliable. For diffusive mixing, model predictions indicate robustness to variation in visual clarity, maintaining high confidence even under low-contrast conditions. This suggests that the class exhibits strong, contrast-independent visual cues, or that the model has learned effective internal representations for this pattern. For transitional mixing, model predictions are moderately clarity-dependent, with greater confidence for higher-contrast droplets. However, a relatively large dispersion of confidence, including at high contrast, suggests visual clarity alone is not responsible for model uncertainty with this class, and that inherent visual inter-class similarity may contribute.

4 Stage 2 - Temporal Classification Network

4.1 Methods

For the second stage of our detection and classification pipeline, we introduce a powerful hybrid network divided into two modules; spatial information processing and temporal information processing. The former comprises a state-of-the-art convolutional neural network (CNN) for spatial feature extraction, enhanced by multiscale feature aggregation and channel-spatial attention submodules. The latter comprises a bidirectional long short-term memory network (BiLSTM) for long-term temporal modelling, enhanced by a temporal attention submodule (TAM), followed with a final classifier. In total, the network has 219 layers and 40,494,437 trainable parameters. Our proposed CNN-BiLSTM-TAM is designed to perform sequence-level classification of transcritical mixing regimes, using the object-level droplet images generated by the first stage of our pipeline, described in Section 3.

Figure 6 is a network schematic illustrating each of these components. For spatial feature extraction, we use the residual split-attention network ResNeSt [12], which improves upon the ResNet [13] architecture through split-attention blocks, which allows the model to learn channel-wise attention within groups. This combines the strengths of multi-branch processing introduced by ResNeXt [14] and dynamic attention

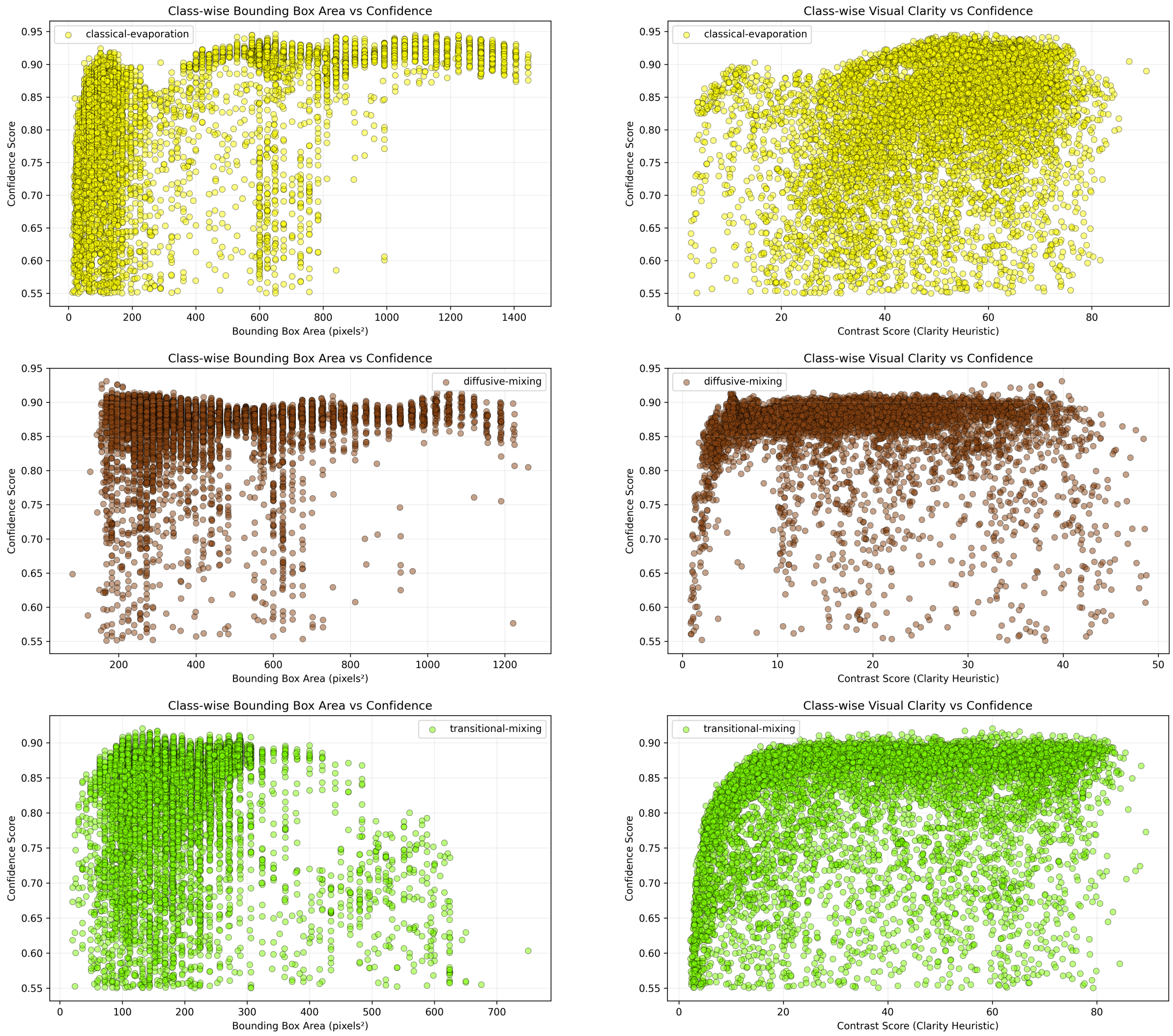


Figure 5. *Class-wise bounding box area and visual clarity heuristic vs YOLOv11 model confidence.*

by SENet [15], improving spatial representation. We use the pretrained (on ImageNet) ResNeSt-50d, enhanced with the ResNet-D tweak introduced in [16], 50 layers deep. We found that this was sufficient for extracting rich and diverse per-frame spatial features, producing feature maps with 2048 channels.

To further enhance spatial context, the spatial pyramid pooling fast (SPPF) module performs multiple levels of max pooling and concatenation to fuse local and global context at multiple receptive fields. The output maintains the same spatial resolution but captures richer information, through scale-invariance and enhanced context awareness. SPPF is a faster, simpler implementation over the original SPPNet [17]. These rich feature maps from SPPF are further refined by a lightweight implementation of the convolutional block attention module (CBAM) [18], which sequentially applies channel attention and spatial attention, encouraging the model to focus on important features and regions. CBAM has shown to be more effective than other attention mechanisms, as it determines both ‘which’ features are important across all spatial locations, and ‘where’ the important features are within the spatial dimensions [18]. It also uses a combination of max pooling and average pooling, where previous architectures like SENet have only used average pooling for channel attention, omitting max pooling and spatial attention entirely [18]. CBAM also outperforms concurrent attention modules like CSSE [19], learning more nuanced features due to its sequential refinement of attention across dimensions. Overall, the use of SPPF and CBAM enhances the discriminative power of the spatial information processing of our network. Lastly, we use global average pooling to compress the refined spatial feature maps into vectors, and then use a fully connected (FC) layer to reduce the channels from 2048 to 1024, a more manageable size for the temporal information processing of our network, improving training speed and efficiency. This layer acts as a bottleneck to distil global spatial semantics into compact vectors for temporal modelling.

The reduced per-frame feature vectors are reshaped into sequences and passed to a BiLSTM [20]. As

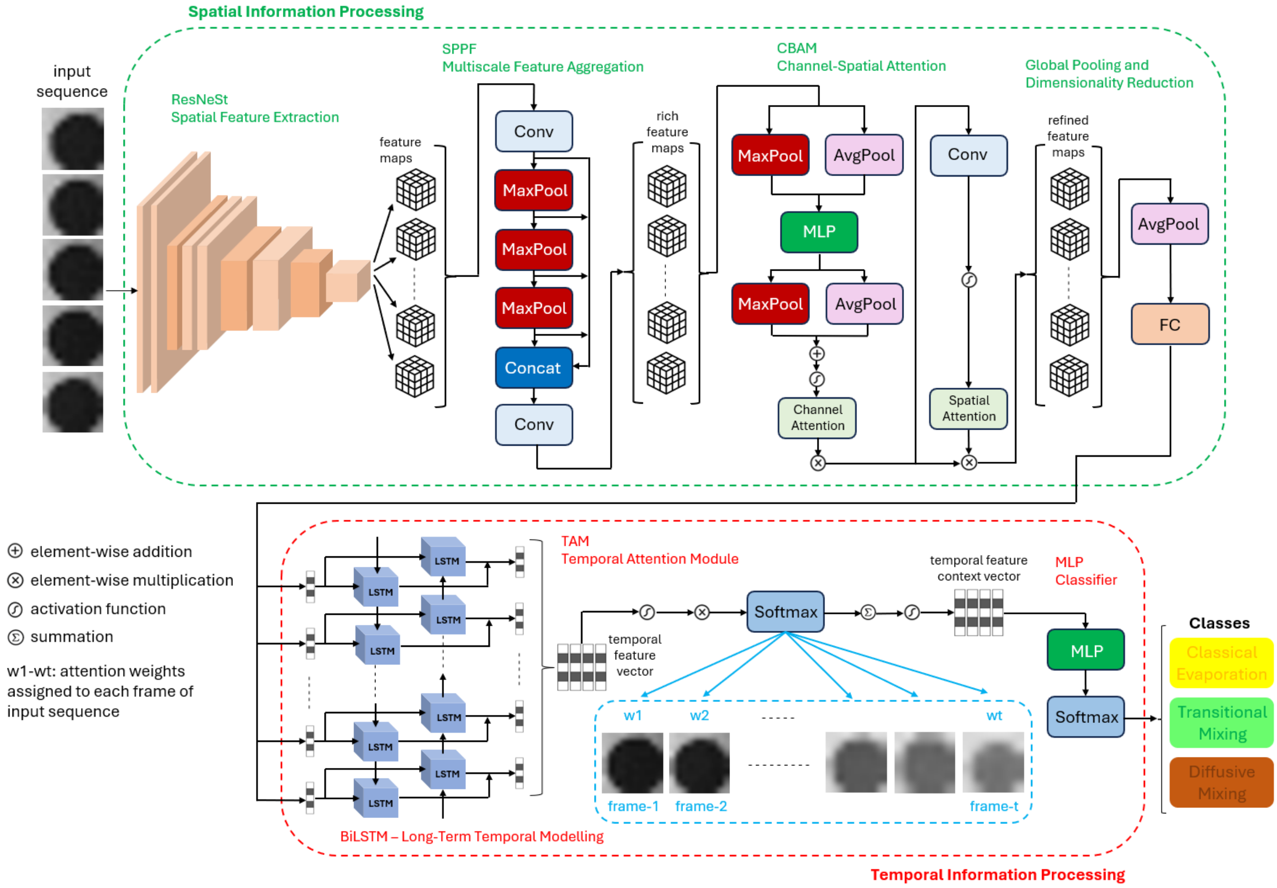


Figure 6. *A detailed schematic of the proposed CNN-BiLSTM-TAM hybrid network, divided into the spatial information processing and temporal information processing modules.*

the length of droplet sequences can vary in maximum frames, from as little as 21 frames to as many as 916, a collate function is employed to group variable-length sequences into padded tensors. We then use packed sequences to allow the BiLSTM to ignore padding, only processing real frames. Due to GPU VRAM constraints, we only use a maximum of 250 frames as inputs to our model. If the sequence has $\leq$250 frames, we use them all, otherwise we use the following methodology: we skip the first 10 frames (as these often show occluded or obscured droplets as they exit the injector region), then take the next 30 frames, ensuring the initial droplet shape is included. We also take the last 90 frames, as the three mixing regimes show more morphological variation towards the end of their lifecycles, and thus this can be highly discriminative information. The middle region therefore consists of 120 frames, which we uniformly sample from the remaining frames to get an accurate view of their temporal evolution.

Due to the sequences being on the object-level, the BiLSTM does not capture motion, but focuses on morphological evolution. We found a hidden size of 192 (in practise, 384 due to the bidirectionality of the LSTM) and a single layer to be sufficient for modelling long-term temporal dependencies. The BiLSTM processes sequences in both forward and backward directions, enabling the model to capture both past and future dependencies. Next, the temporal attention module (TAM) learns to assign weights to each frame in the sequence, focusing on the most informative ones for classification [21]. Firstly, we compute a per-timestep energy score using a trainable parameter vector. Then the scores are passed through a softmax function to generate attention weights over the time steps, and then the BiLSTM outputs are weighted accordingly and summed across time to form a context vector [21]. As we are using padded sequences, masks are used in TAM to exclude padded frames from the weighted sum. This context vector captures the most informative moments in the sequence, reducing noise from redundant frames, thus improving performance where no temporal attention has the BiLSTM treat all frames equally.

The TAM context vector is then fed through a multilayer perceptron (MLP) to produce the final class logits. The MLP consists of two FC layers, with ReLU activation function for non-linearity and dropout to regularise the network in-between. Softmax then transforms the logits into a probability distribution where each output represents the likelihood of belonging to a specific transcritical mixing regime class.

4.2 Results

The object-level dataset consists of 103 droplet sequences, with a total of 26,939 frames. The dataset is split into train, validation, and test sets via a 80/10/10 split that maintains relative class balance across sets. All frames use a spatial resolution of 224x224. Sequence consistent augmentation is applied to all droplet sequences in the train set. This applies augmentation procedures across all frames of a sequence to maintain visual-temporal consistency. We use resized cropping to simulate zoom, horizontal flip to encourage left-right invariance, colour jitter (brightness, contrast, and saturation) to handle lighting variations, and Gaussian blur to smooth sharp transitions.

Model training took 2.5 hours for a maximum of 100 epochs on a RTX 3090 Ti GPU. Due to a maximum of 250 frames per sequence as model input, a batch size of 1 was selected to work with the 24 GB VRAM. We utilise decoupled learning rates (LR), applying a higher LR of 1e-4 to the BiLSTM and MLP, and a smaller LR of 1e-5 to the pretrained CNN backbone, to avoid destroying learned ImageNet features. In addition, we freeze the early layers of the CNN backbone initially for 7 epochs, preserving low-level spatial features from pretraining. To stabilise training, we gradually increase the LR from a small value to the base LR over the first 3 epochs, then use cosine annealing with periodical restarts for the LR in a cosine curve pattern every 10 epochs. We use the AdamP optimiser [22], to prevent over-adaptive updates, and label smoothing to reduce model overconfidence, overall improving generalisation. We apply L2 regularisation to the weights with 1e-4 decay during training and early-stopping after 20 epochs, to help prevent overfitting on the train set. Where possible, we speed up training using mixed precision.

Table 3 summarises the results of our hybrid network on the test set. As shown in Table 3, our hybrid network is consistently accurate and robust for complex sequence-level classification of transcritical mixing regimes, utilising both spatial and temporal context to inform classification. This can be observed when compared to our YOLOv5 model from [2], where our CNN-BiLSTM-TAM improves precision, recall, and F1 score by +48%, +38%, and +42% respectively.

Table 3. *Classification results of our CNN-BiLSTM-TAM on the test dataset compared to SOTA. The F1 score is the harmonic mean of the precision and recall. The AUC is the area under the curve, where the curve is the receiver-operating characteristic (ROC). The MCC is Matthews correlation coefficient, in this context a measure of the quality of multi-class classifications. The value lies between +1 and -1, where +1 is a perfect prediction, 0 an average random prediction, and -1 an inverse prediction.*

	Accuracy	Precision	Recall	F1	AUC	MCC
YOLOv5 [2]	-	0.45	0.51	0.48	-	-
YOLOv11 [6]	0.64	0.43	0.58	0.50	0.55	0.48
YOLOv12 [23]	0.36	0.12	0.33	0.18	0.84	0.00
CNN-BiLSTM-TAM	**0.91**	**0.93**	**0.89**	**0.90**	**0.90**	**0.87**

Direct comparison between our CNN-BiLSTM-TAM and YOLOv5 is not ideal, as our YOLOv5 is an older architecture, trained and validated on a different dataset [2]. For a fairer comparison, we train two state-of-the-art (SOTA) image classifiers on our synthetic droplet dataset; YOLOv11 [6] and YOLOv12 [23]. For both SOTA models we used the medium scale architecture with no modifications aside from stripping the final classification head layer. As these classifiers cannot model temporality, we obtained sequence-level predictions by performing frame-wise classification on each frame in the sequence, then computing the average softmax probability for the final prediction. Each SOTA model was trained for a maximum of 100 epochs on a RTX 3090 Ti GPU, maintaining the same 250 frames-per-sequence and batch size of 1 training strategy as our network. A LR of 1e-5 and the AdamW [11] optimiser was used as these are YOLO standards. We used cosine annealing and early stopping after 20 epochs to prevent overfitting.

As explored in Section 1, conventional ML models often fail to resolve the time-variant features required for accurate classification of arbitrary transcritical droplets. With our previous models, YOLOv5 and Mask R-CNN, we observed a similar ceiling in performance regardless of the quality of our model architecture or training dataset. We hypothesised that omitting the temporal context, as had been done in previous studies, was degrading the performance of our classification algorithms by not exposing the full spatiotemporal patterns to be learned. This can be observed in Table 3, where our CNN-BiLSTM-TAM outperforms the SOTA models by a considerable margin. Compared to YOLOv11, our CNN-BiLSTM-TAM improves precision, recall, F1 score, and MCC by +50%, +31%, +40%, and +39% respectively, whereas with YOLOv12 it improves these metrics by +81%, +56%, +72%, and +87% respectively. Notably, YOLOv12 performed worse than YOLOv11 despite being shown to outperform it [23]. We speculate that this is likely due to performance benefit from transfer learning, as YOLOv11 was initialised with pretrained weights whereas such weights were not available for YOLOv12 at the time of writing. Based on the results of Table 3, we can conclude that leveraging full spatiotemporal patterns can better inform complex sequence-level classification of transcritical mixing regimes.

A benefit of our model by design is its interpretable nature. To provide insight into its class-specific focus patterns, we visualise the spatial and temporal attention maps for representative test sequences of all three classes. In Figures 7, 9, and 10, the 1st row represents the ground truth frame, the 2nd the spatial attention map from the ResNeSt backbone using gradient class activation mapping (grad-CAM), the 3rd the spatial attention map from the CBAM submodule after SPPF extracted directly from the sigmoid weights, and the 4th an overlay of alpha weights from the TAM submodule indicating temporal attention. In the bottom right corner is the predicted class label and softmax score as model confidence.

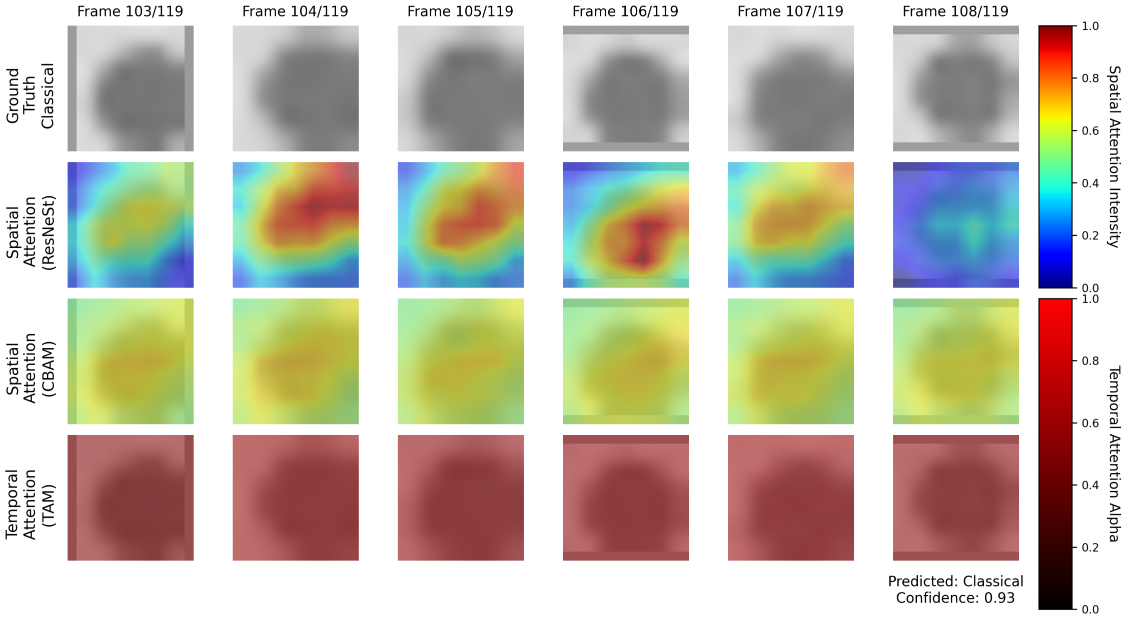


Figure 7. *Spatial and temporal attention maps of the CNN-BiLSTM-TAM model for a 6-frame slice towards the end of a classical regime sequence correctly predicted with high confidence.*

In Figure 7, we can observe that the ResNeSt spatial attention is highly concentrated around the droplet's central body, with vivid red focal regions indicating strong gradient-based importance in relation to the shape and contrast features. The CBAM spatial attention is more diffuse, with a consistent but milder attention across the entire droplet region. This reinforces broad focus on the droplet shape, potentially capturing fine-grained texture features. For TAM temporal attention, the alpha weights are high and consistent, indicating that these frames are considered highly informative for sequence-level classification.

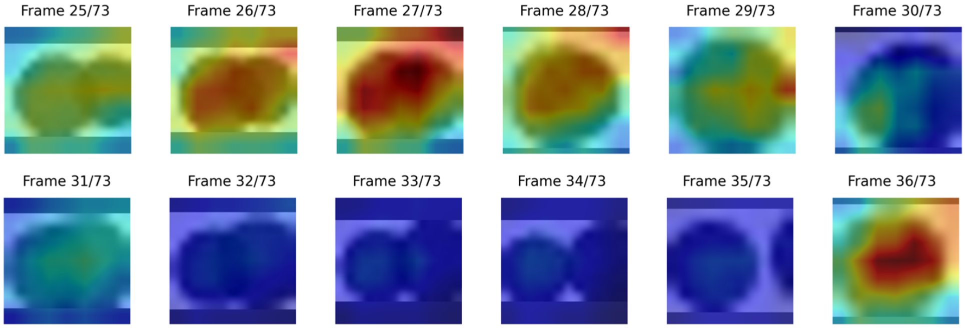


Figure 8. *Spatial attention maps of the CNN-BiLSTM-TAM model for a 12-frame slice in the middle of a classical regime sequence, demonstrating robustness to occlusion via rogue droplets.*

In some cases, occlusion would occur where rogue droplets would be present in frame; one such example can be seen in Figure 8. Interestingly, the initial frames show some ResNeSt spatial attention activation, particularly on frame 27. Later in the sequence (frames 32-35), there is a distinct lack of activation, which begins to pick up again in frame 36 where the droplet is no longer occluded. We speculate that the ResNeSt network has learned to ignore occluded frames interspersed in droplet sequences.

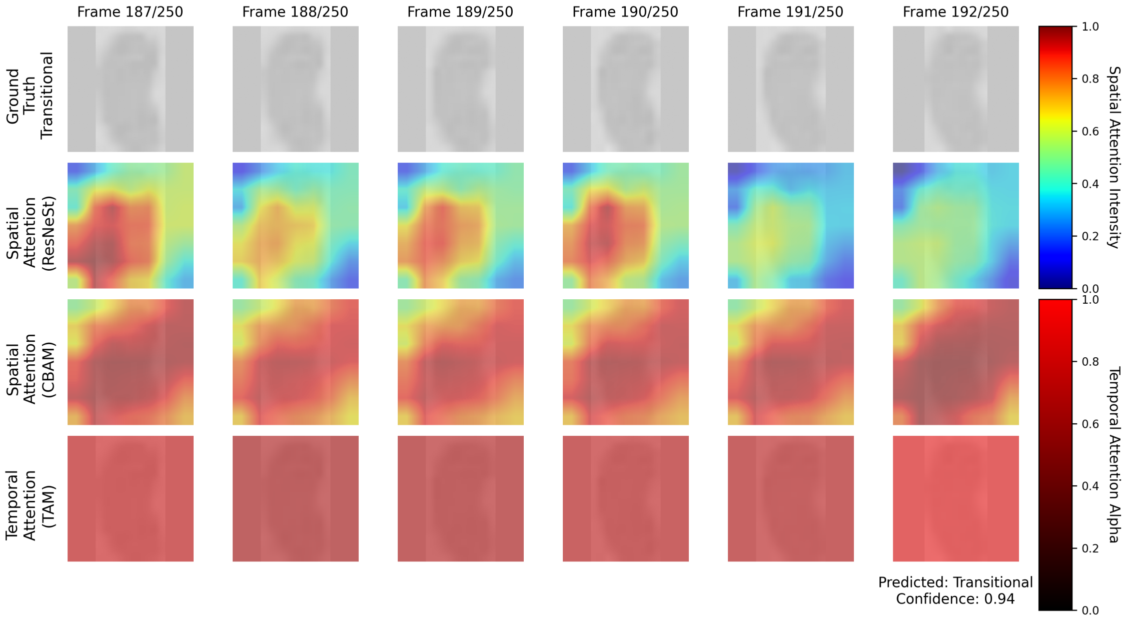


Figure 9. *Spatial and temporal attention maps of the CNN-BiLSTM-TAM model for a 6-frame slice towards the end of a transitional regime sequence correctly predicted with high confidence.*

In Figure 9, we can observe that the ResNeSt spatial attention is highly concentrated around the middle-left zone, with spatial focus shifting subtly across frames, tracking changes in droplet contours. The CBAM spatial attention appears to consistently agree with ResNeSt in high-activation zones across frames, with broad attention almost saturating across the full droplet region. Slight gradient in intensity from lower-left to upper-right suggests prioritisation of certain curvature regions. For TAM temporal attention, very high alpha weights signify that these frames contain critical temporal cues for robust identification of this class, and are heavily weighted in the sequence-level classification.

In Figure 10, we can observe that the ResNeSt spatial attention exhibits a clear increase in spatial selectivity toward the lower-right region of the droplet, initially weaker and diffuse for earlier frames, then intensifying, learning to focus on subtle structural features that may be evident of deformations associated with the diffusive regime. The CBAM spatial attention has uniformly mild attention across frames, with slightly enhanced weighting near the central region of the droplet's body. For TAM temporal attention, the alpha weights are initially low, but towards the end of the sequence gradually increasing, indicating that only the final few frames highly informative for sequence-level classification.

5 Conclusions

In this paper, we demonstrate a two-stage detection and classification pipeline, where the state-of-the-art object detector YOLOv11 and integrated tracker BoT-SORT detect moving droplets and isolate them at the droplet level. YOLOv11 and BoT-SORT outperformed our previous deep learning models and achieved a precision, recall, and mAP@50-95 of 1, 0.81, and 0.79 respectively. At a confidence threshold of 55%, a mean F1 of 0.88 was observed. For the diffusive regime, YOLOv11 demonstrated robust performance regardless of droplet scale and visual clarity, whereas for the classical regime, the opposite was observed. For the transitional regime, detection performance was closer to the diffusive regime, though smaller and lower-contrast cases proved as difficult as the classical regime.

Next, a powerful hybrid network CNN-BiLSTM-TAM is proposed to perform sequence-level classification, utilising both the spatial and temporal characteristics of the droplets to inform classification, achieving an F1 score of 0.90, AUC of 0.90, and MCC of 0.87. This indicates a strong and robust understanding of the visual properties and morphological behaviours of all three transcritical droplet regimes. This was further explored through analysis of the spatial and temporal attention maps, where it can be concluded that

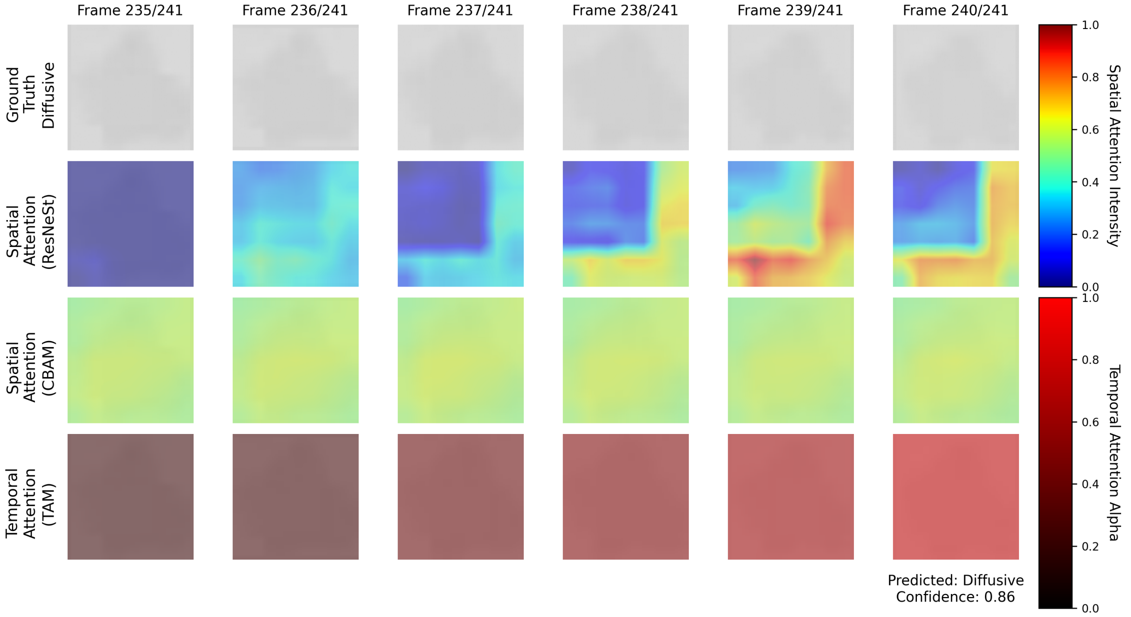


Figure 10. *Spatial and temporal attention maps of the CNN-BiLSTM-TAM model for a 6-frame slice towards the start of a diffusive regime sequence correctly predicted with relatively high confidence.*

ResNeSt backbone is able to learn to isolate complex class-specific spatial features, and interestingly, shows a lack of focus for occluded frames interspersed within test sequences. The CBAM submodule acts as a stabilising spatial mechanism that supports global context awareness, as it uses rich feature maps from the SPPF submodule. CBAM is able to reinforce spatial salience across frames (in the case of the transitional regime) or reinforce broad focus and balance noisy ResNeSt attention (in the case of the diffusive regime). TAM is able to evaluate which frames in the sequence are most relevant for sequence-level classification. Interestingly, we observed that frames towards the end of a droplet's lifecycle are often weighted more heavily by CBAM and TAM, indicating that the class-wise differences in droplet morphology are far more discriminative at this point in the sequence than any other. The exception to this is the diffusive regime, which seemed to have very little temporal weighting save for the final 3-5 frames. This suggests that spatial features were more informative in classifying this regime than temporal context, likely due to the variation in visual deformities.

Future work will develop our pipeline further, improving the multi-scale prediction architecture of YOLOv11 to better detect small, fast-moving, and low-contrast droplets, particularly of the classical and transitional regimes. Better prediction will enable automated generation of object-level droplet sequence datasets without human oversight. We will expand the temporal context window of our CNN-BiLSTM-TAM model, to enable better long-term temporal dependency modelling of longer droplet sequences, with emphasis on latter regions of the sequence. Our pipeline will be validated on real world datasets from [1, 2].

References

[1] C. Crua, J. Manin, and L. M. Pickett, "On the transcritical mixing of fuels at diesel engine conditions," Fuel **208**, 535–548 (2017).

[2] F. Di Sabatino, K. Wan, J. Manin, T. Capil, Y. Hicks, A. Gander, and C. Crua, "The role of diffusive mixing in current and future aviation fuels at relevant operating conditions," Journal of Engineering for Gas Turbines and Power **146** (2024).

[3] G. Jocher, "Ultralytics yolov5," `https://docs.ultralytics.com/yolov5/` (2020). Accessed: 17/3/2025.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision,* (2017), pp. 2961–2969.

[5] C. Doche, "Fuel spray imaging and analysis using machine learning," Tech. rep., Advanced Engineering Centre, University of Brighton (2023).

[6] P. Hidayatullah, N. Syakrani, M. R. Sholahuddin, T. Gelar, and R. Tubagus, "Yolov8 to yolo11: A comprehensive architecture in-depth comparative review," arXiv preprint arXiv:2501.13400 (2025).

[7] N. Jegham, C. Y. Koh, M. Abdelatti, and A. Hendawi, "Evaluating the evolution of yolo (you only look once) models: A comprehensive benchmark study of yolo11 and its predecessors," arXiv preprint arXiv:2411.00201 (2024).

[8] J. Bergstra and B. Kégl, "Algorithms for hyper-parameter optimization," Advances in Neural Information Processing Systems **24** (2011).

[9] L. Li, K. Jamieson, A. Rostamizadeh, E. Gonina, J. Ben-Tzur, M. Hardt, B. Recht, and A. Talwalkar, "A system for massively parallel hyperparameter tuning," Proceedings of machine learning and systems **2**, 230–246 (2020).

[10] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "Bot-sort: Robust associations multi-pedestrian tracking," arXiv preprint arXiv:2206.14651 (2022).

[11] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations,* (2019).

[12] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* (2022), pp. 2736–2746.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* (2016), pp. 770–778.

[14] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* (2017), pp. 1492–1500.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* (2018), pp. 7132–7141.

[16] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* (2019), pp. 558–567.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE transactions on pattern analysis and machine intelligence **37**, 1904–1916 (2015).

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV),* (2018), pp. 3–19.

[19] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I,* (Springer, 2018), pp. 421–429.

[20] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," Neural networks **18**, 602–610 (2005).

[21] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers),* (2016), pp. 207–212.

[22] B. Heo, S. Chun, S. J. Oh, D. Han, S. Yun, G. Kim, Y. Uh, and J.-W. Ha, "Adamp: Slowing down the slowdown for momentum optimizers on scale-invariant weights," arXiv preprint arXiv:2006.08217 (2020).

[23] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," arXiv preprint arXiv:2502.12524 (2025).