

Goal and overview

Goal

Adapt a vision transformer model with temporal memory to track every droplet in high-speed videography of GMAW and quantify detachment frequency, droplet radius, volume, and acceleration.

Overview

This study adapts EfficientTAM, a vision transformer model whose temporal memory keeps one identity per droplet.

Background

Welding performance depends on current and voltage, whose effect on metal transfer guides power source design. High-speed videography enables that study, but manual measurement does not scale, and frame-by-frame deep learning loses the droplet identity. Vision transformers overcome both limitations, since their temporal memory keeps one identity per droplet. Of the five regions in Fig. 1 (Alvarez-Rocha et al., 2026), only the droplet and the molten consumable are measured, while the internal and external arc segmentations are not used.

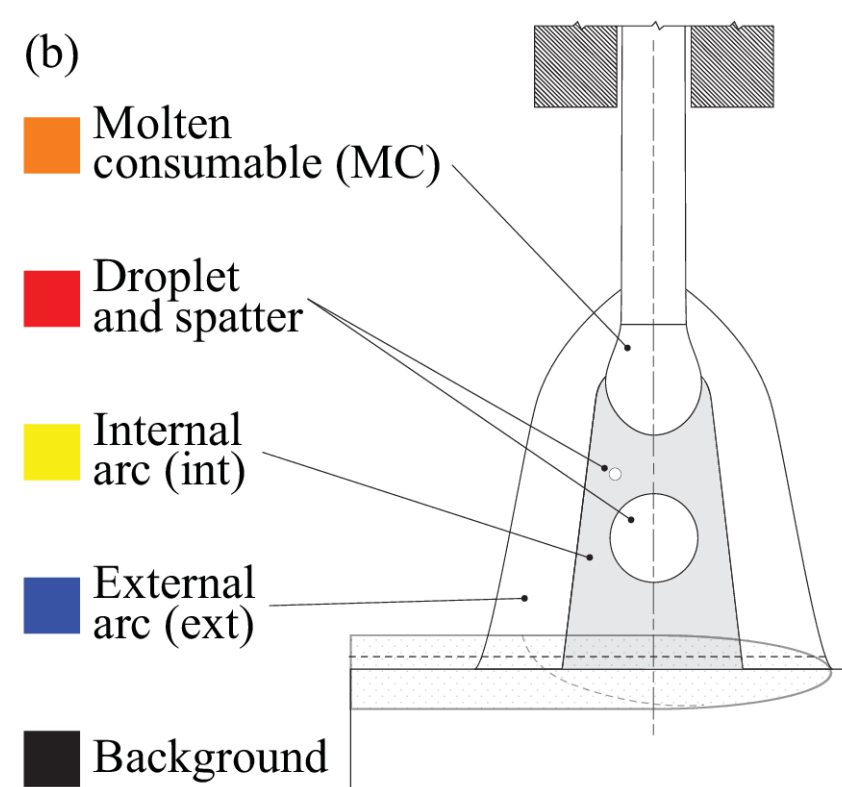


Fig. 1: Frame, arc regions and manual mask.

Methods

Dataset

The dataset comprises 24 high-speed videos of about 15,000 frames each, GMAW with 1.6 mm ER4043 aluminium wire, covering 122 to 204 ipm and 21 to 29 V. A total of 1,204 frames were manually extracted and annotated (Alvarez-Rocha et al., 2026), the source of the pixel-to-mm scale.

- **Train** (800) teaches the model
- **Validation** (200) tunes the training choices
- **Test** (204) stays outside the training range

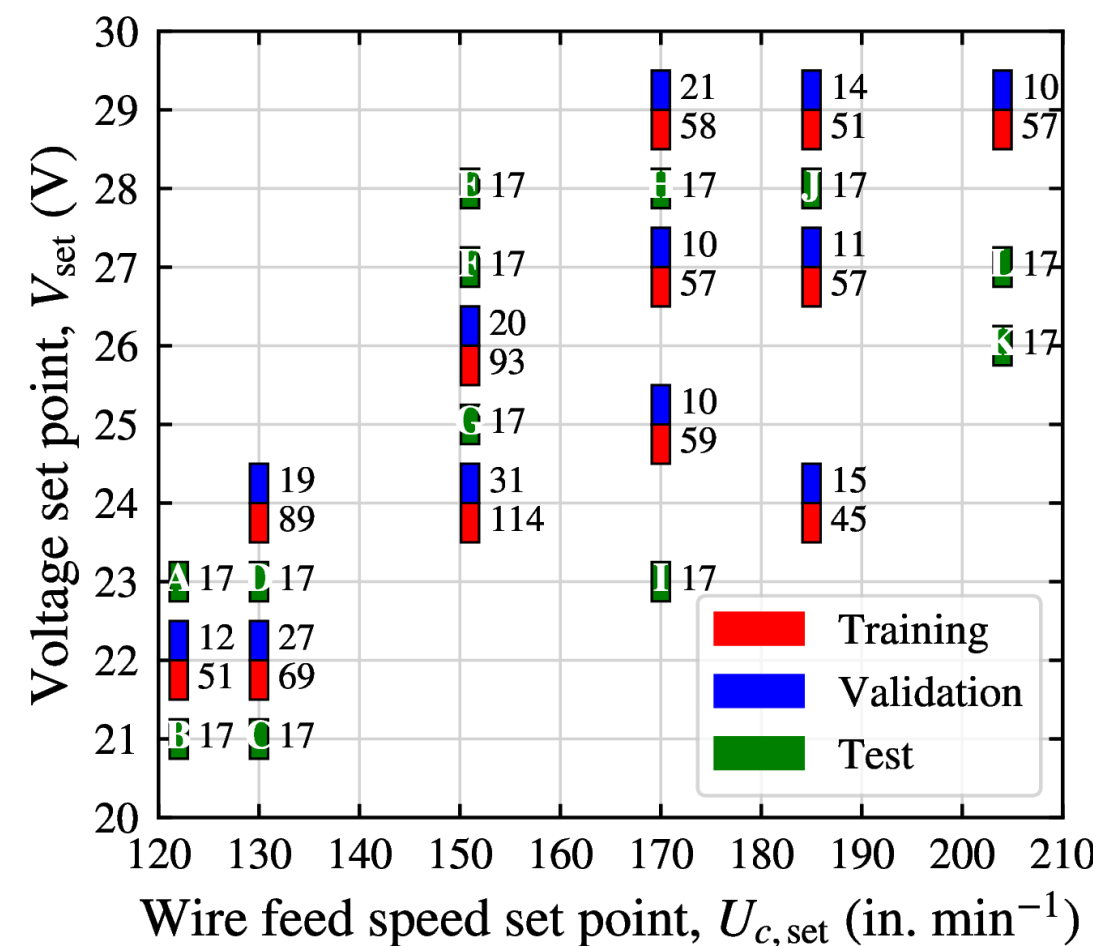


Fig. 2: Data split across welding conditions.

Adapting EfficientTAM to welding

EfficientTAM (ETAM) integrates a lightweight vision transformer with an efficient temporal memory. It works from prompts, boxes or points that indicate the object, and from a single prompt the memory tracks that droplet across the whole video. The frozen model was adapted with two modifications (Fig. 3), where Phase 1 adds automatic five-class segmentation and prompts, and Phase 2 adds a refiner that corrects the droplet masks. The droplet IoU (intersection over union, 1 is perfect overlap) rises from 0.761 with box prompts to 0.862 with the full chain (Table 1).

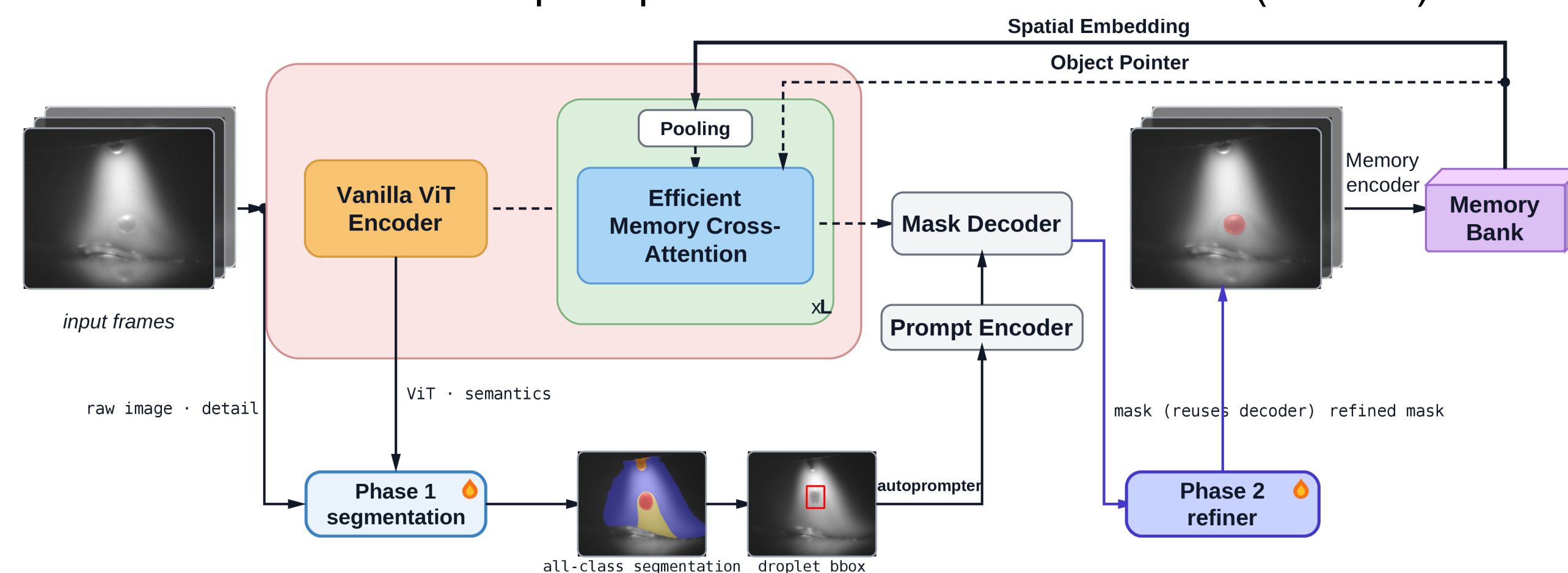


Fig. 3: The two adaptations, Phase 1 (segmentation and prompts) and Phase 2 (refiner). Base architecture adapted from Xiong et al. (2024).

	IoU	
Stage	Droplet	Molten consumable
ETAM using box prompts*	0.761	0.638
Phase 1: segmentation	0.839	0.747
Phase 2: refiner	0.862	0.737

Table 1: Semantic IoU on the held-out test set. *Earlier protocol, scored per droplet with manual boxes on an earlier ground-truth revision, so not directly comparable.

Segmentation and tracking system

The system runs segment, seed and track, then validates each trajectory (Fig. 4), with memory attention carrying each identity across frames.

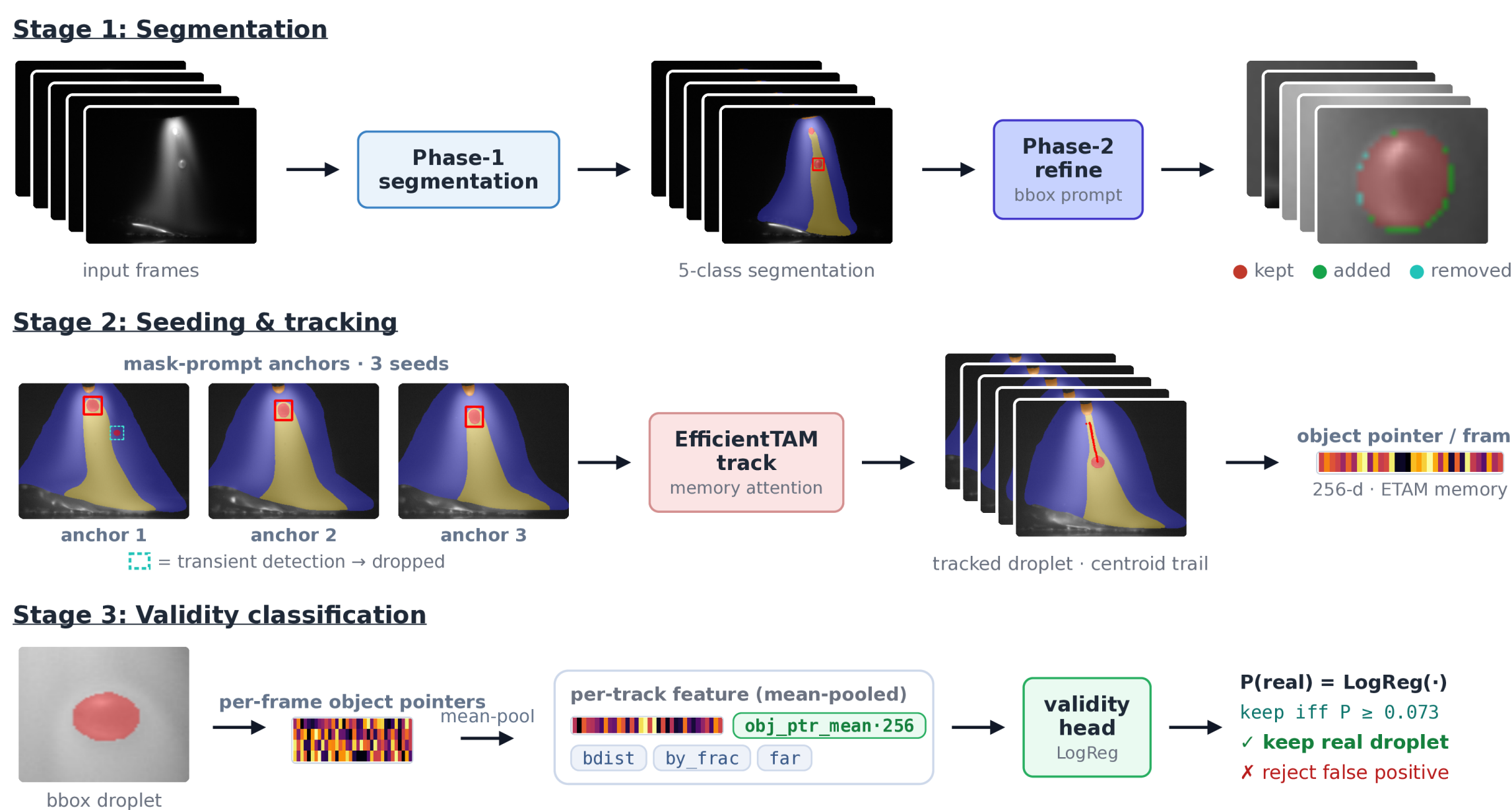


Fig. 4: Staged segmentation-and-tracking pipeline. Inspired by Amini-Naieni & Zisserman (2025).

Processing the high-speed video

All measurements come from the droplet trajectories delivered by the system. Since the 15,000 frames of a video do not fit in GPU memory, it is processed in overlapping windows stitched in the overlap, keeping one identity per droplet.

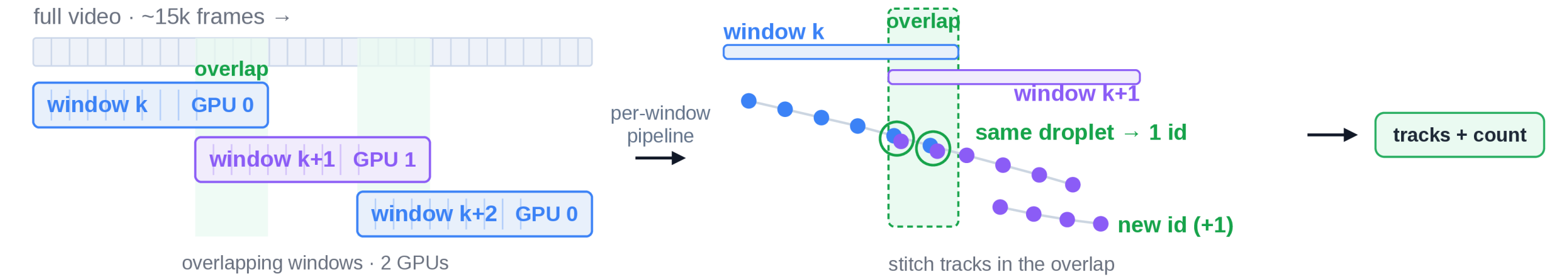


Fig. 5: Overlapping windows keep one identity per droplet.

For each droplet, the projected area gives an equivalent radius and the volume is the sphere of the mean radius, while the reference volume comes from a mass balance of the fed wire, using the wire diameter and the feed speed (Fig. 6).

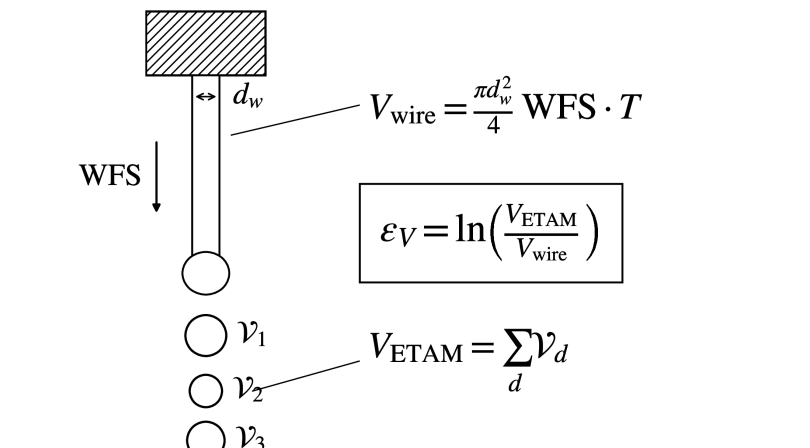


Fig. 6: Volume error estimate.

Results

Each point is one of the 24 videos. The frequency error $\epsilon_f = \ln(f_{ETAM}/f_{manual})$ stays below 4 % in every video, with an exact manual-count match in 7 of them. Only accepted trajectories are counted, so spatter is excluded.

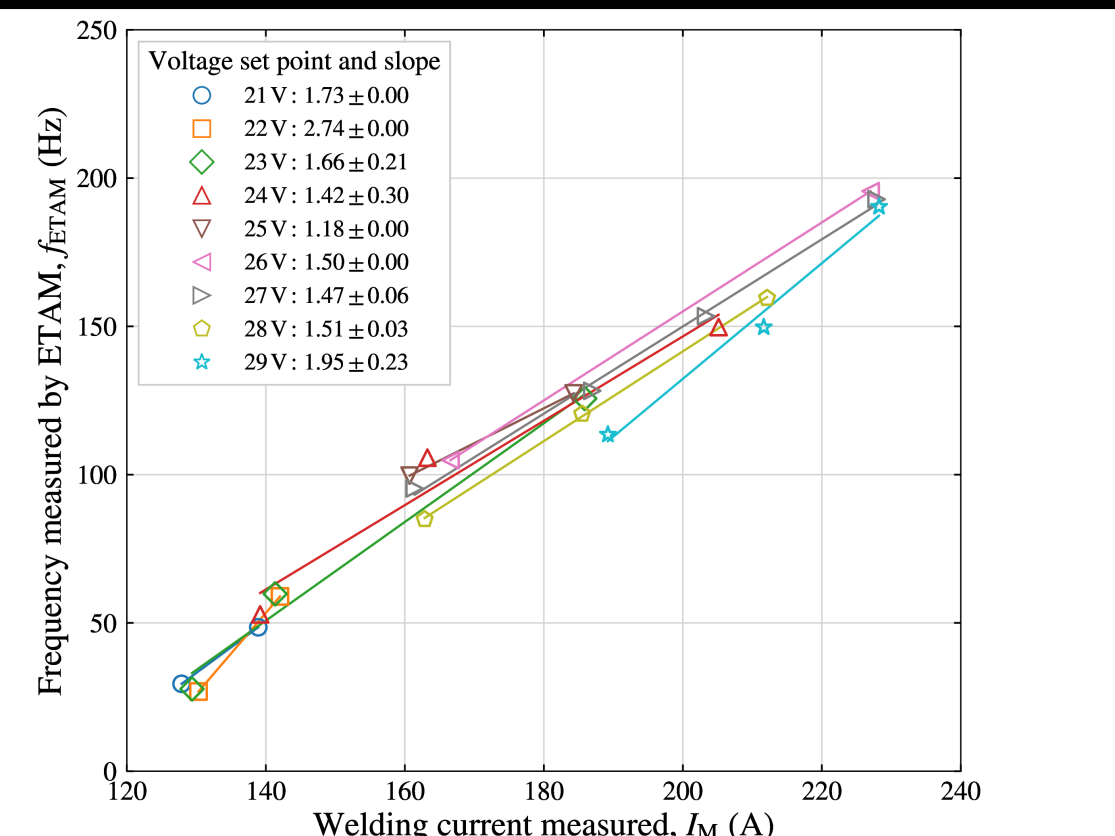


Fig. 7: Transfer frequency increases with current at every voltage set point.

The mean radius decreases with current, with an increase in the boxed region above 200 A (Fig. 8). Sizing every droplet individually resolves the dispersion within each condition, where two average sizes appear (Fig. 9, bimodal behavior).

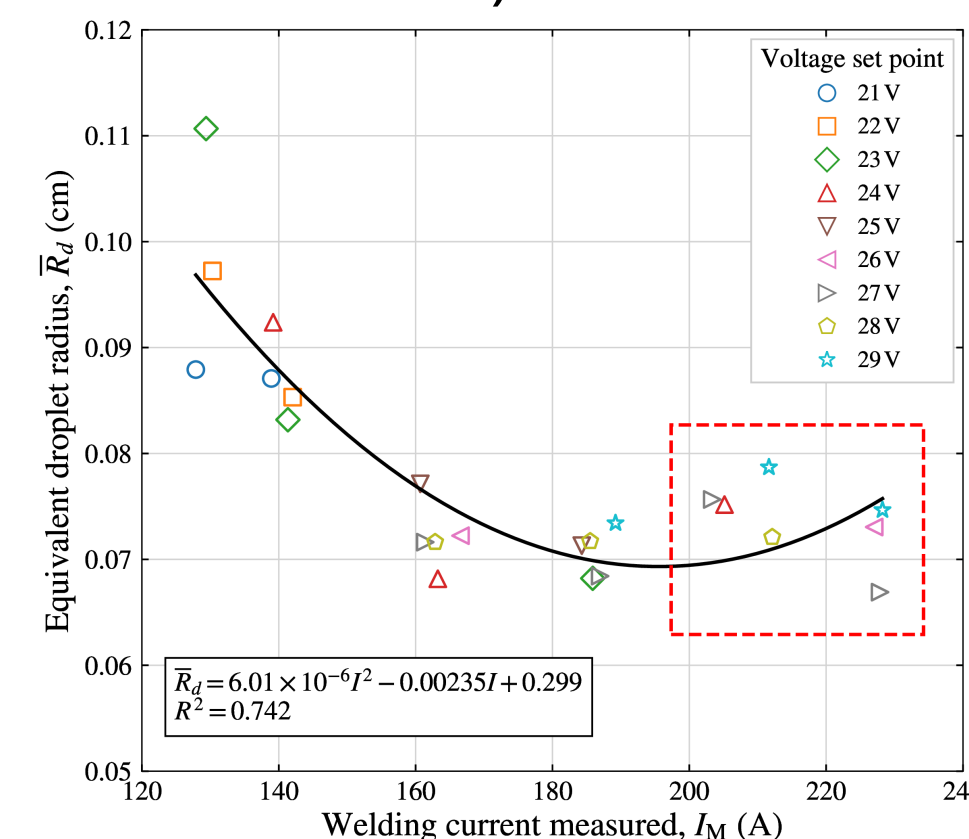


Fig. 8: Mean droplet radius per video vs current.

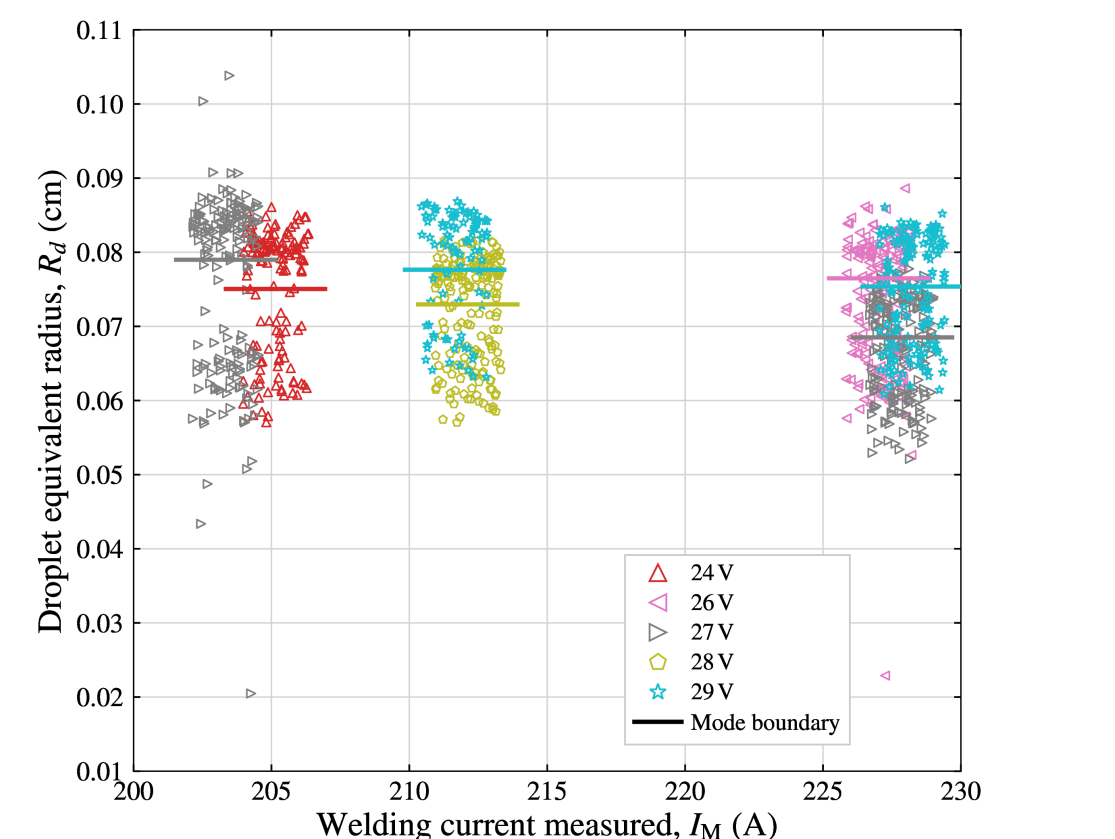


Fig. 9: One marker per droplet (95 to 184 per video) above 200 A.

The measured volume deviates 46 % from the wire-fed reference (Fig. 6), a systematic overestimate. The plasma drag force was not estimated, since this uncertainty would dominate its error, and the reported acceleration grows with current far beyond gravity (Fig. 10).

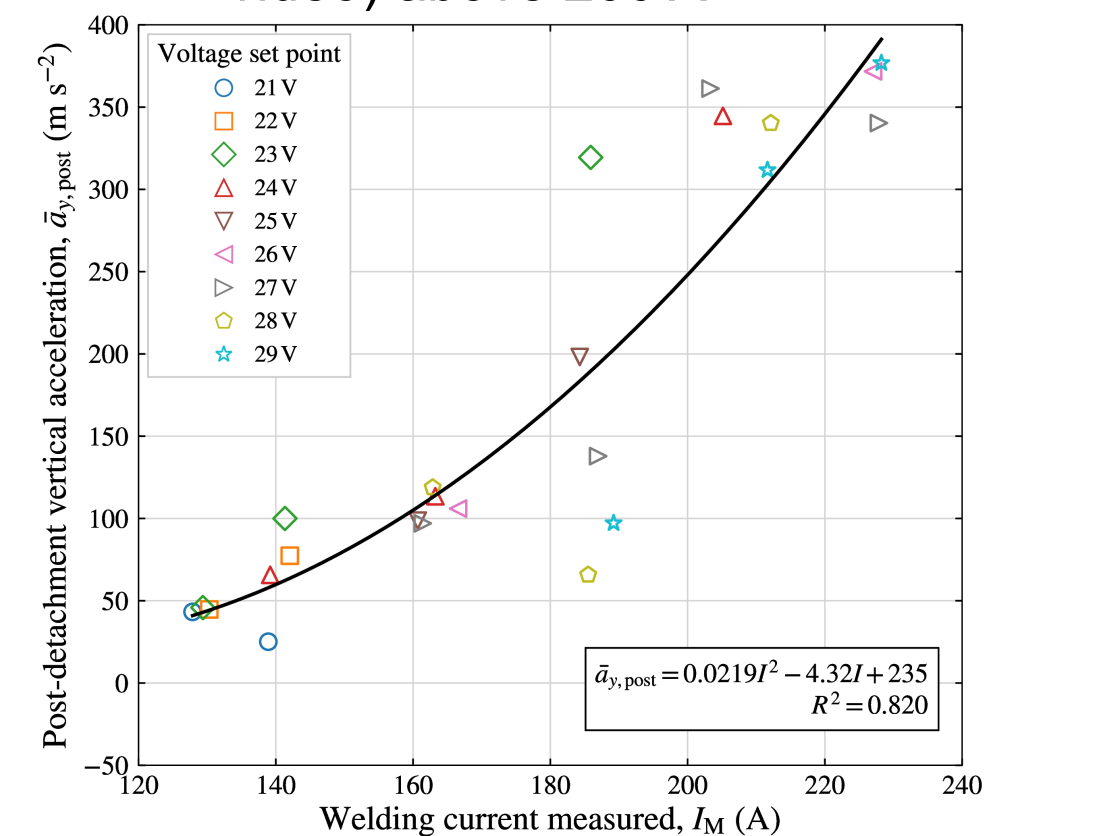


Fig. 10: Mean vertical acceleration per video, quadratic fit over the initial 30 % of the fall.

Discussion

- Above 200 A two average droplet sizes coexist (Fig. 9), which explains the rise of the mean radius in Fig. 8
- Volume scales with the cube of the radius, so a 1 to 2 px contour error strongly affects it, especially for small droplets
- Temporal memory improves segmentation continuity during droplet occlusions and overlap

Conclusion and future work

- A droplet IoU of 0.862 and a frequency error below 4 % per video (1.25 % on average) demonstrate reliable automated frequency measurements
- 25 to 30 minutes per video on two RTX A4500 GPUs make fully automated analysis practical
- Accurate volume estimates require higher resolution or improved backlighting

Future work

- Improve droplet contour definition using laser backlighting
- Estimate plasma drag force after reducing volume uncertainty, which currently dominates the force error

Acknowledgments

This work was supported by Lincoln Electric.
Contact: taiperam@ualberta.ca