A Systematic Evaluation of Deep Learning Architectures and Training Strategies for Multi-Class Raindrop Segmentation

Kento Ichihara[1], Lin Meng[2]
{ri0135sx@ed.ritsumei.ac.jp[1],menglin@fc.ritsumei.ac.jp[2]}

Graduate School of Science and Engineering, Ritsumeikan University,
1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577 Japan[1]
College of Science and Engineering, Ritsumeikan University,
1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577 Japan[2]

Abstract. Raindrops adhering to camera lenses degrade image quality and impair downstream computer vision tasks in applications such as autonomous driving and surveillance. Most rain detection methods treat all rain as a single class. We introduce multi-class raindrop semantic segmentation, defining four morphological classes: large drops, small drops, blurred drops, and streaks. We benchmark segmentation architectures, encoder backbones (CNNs and Vision Transformers), loss functions, and optimizers on a custom-annotated dataset using 5-fold cross-validation. A U-Net with a Mix Transformer (MiT-B4) encoder, trained with combined Cross-Entropy and Dice loss, AdamW, and Cosine Annealing, achieves a mean IoU of 0.613 and serves as a baseline for this task.

Keywords: Computer Vision; Semantic Segmentation; Deep Learning; Raindrop Detection; Ablation Study; Vision Transformer

1 Introduction

In recent years, outdoor camera systems have been deployed widely for autonomous navigation, traffic monitoring, and public safety. Their reliability depends on image quality, but rain introduces visual artifacts that impair downstream computer vision algorithms [1, 2]. Deraining—the removal of such artifacts—is a well-studied problem, but its goal is image restoration, not artifact localization. Our group works on single-image deraining, and we have found that even recent methods fail on large, highly distorting raindrops ($\geq$ 50 pixels in their largest dimension) [9]. These artifacts distort the image structure through complex light refraction and scattering, and they are the main failure mode for both restoration and object detection pipelines. Missing a large drop can cause incorrect detections in safety-critical scenarios such as autonomous driving. To address this,

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365273

we propose multi-class raindrop semantic segmentation as a pre-processing step: by localizing and classifying each artifact, large drops can be flagged for specialized restoration or excluded from detection, improving pipeline robustness. Treating all raindrops as a single class introduces high intra-class variance that makes it hard to learn consistent representations; our multi-class formulation lets each morphological type be learned separately. We evaluate deep learning models and training strategies to find the best configuration for this task.

Our contributions are as follows:

- We release a manually annotated dataset for multi-class raindrop segmentation with four classes—large drops, small drops, blurred drops, and streaks—each motivated by a specific failure mode in deraining (Fig. 1).
- We benchmark four segmentation architectures and five encoder backbones (CNNs and Vision Transformers) together with a full set of training strategies under 5-fold cross-validation.
- We show that U-Net with MiT-B4 reaches an IoU of 0.613 and quantify the contribution of each training choice through ablation.

2 Related Work

2.1 Rain Detection and Removal

The removal of rain effects from images, known as deraining, is a well-established field in computer vision [1]. Deraining aims to restore the image, whereas we aim to precisely locate and classify the rain artifacts themselves.

2.2 Deep Learning for Weather Segmentation

Semantic segmentation has been applied to detect various adverse weather phenomena using architectures like U-Net [3] or DeepLab [4]. These methods typically consolidate all rain artifacts into a single "rain" class. Our work extends this line of research by decomposing this single class into multiple, morphologically distinct sub-classes.

2.3 Vision Transformers in Segmentation

The Vision Transformer (ViT) [5] and hierarchical variants such as MiT [6] have shown strong performance in segmentation by capturing global dependencies through self-attention. The Swin Transformer [10] further demonstrates the benefit of multi-scale feature extraction in Transformer architectures. Reducing the computational cost of these models via token reduction has also attracted attention [11, 12].

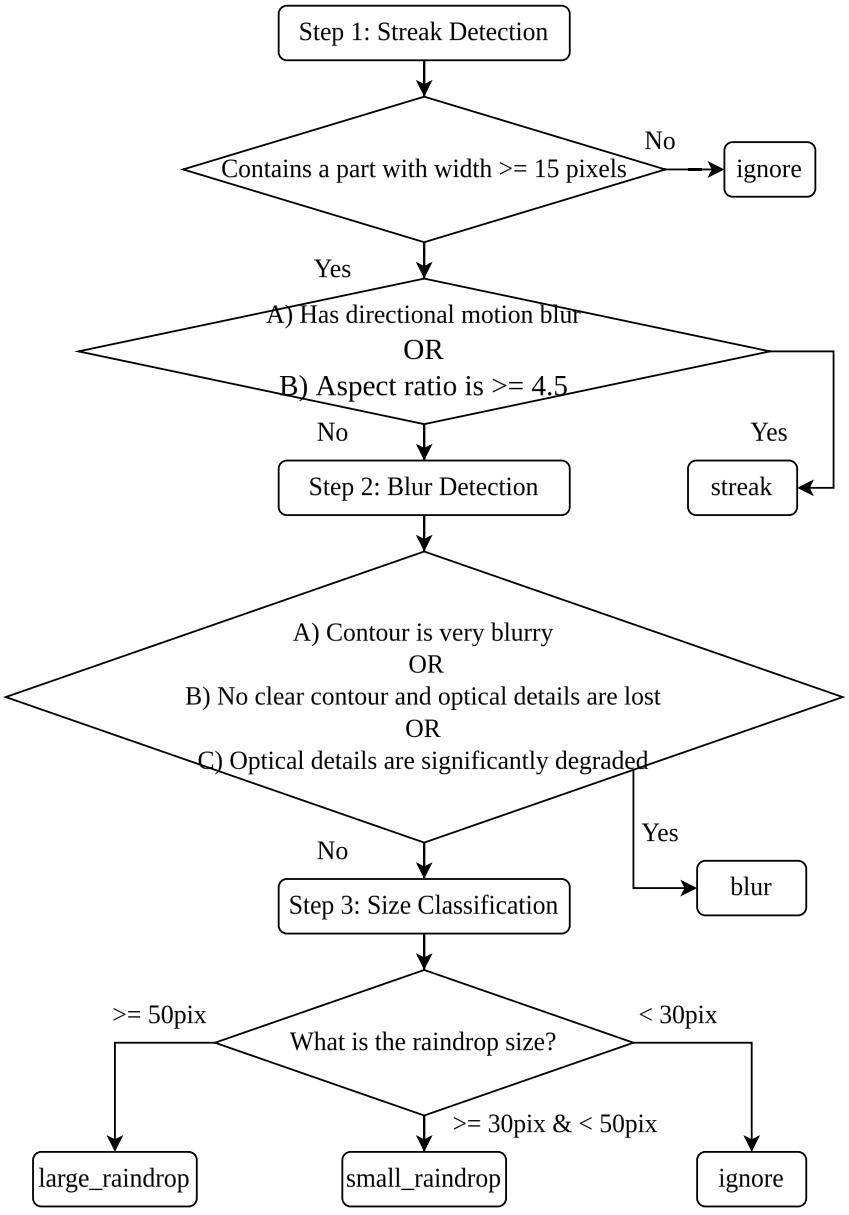


Fig. 1. Annotation Classification Flowchart for Multi-Class Raindrop Segmentation.

3 Experimental Setup

All experiments use 5-fold cross-validation on a single NVIDIA RTX 4090 GPU. The raindrop classes are defined around our target application: handling artifacts that cause existing deraining and object detection models to fail. Large raindrops (≥ 50 pixels) are the hardest for current deraining algorithms, and reliably detecting them is our primary goal. We also include smaller raindrops (≥ 30 pixels) to assess model behavior across a wider size range.

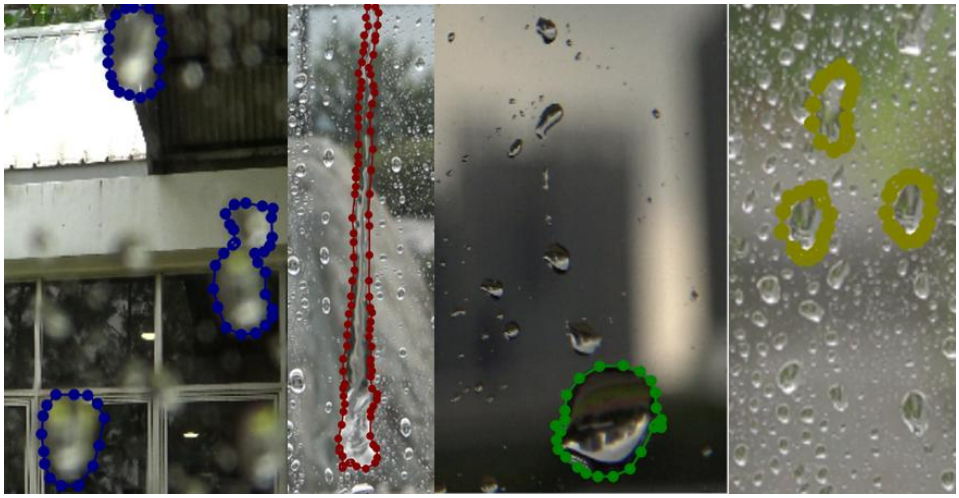


Fig. 2. Visual Examples of the Four Morphological Raindrop Classes. Arranged from left to right: Blur, Streak, Large Drop, and Small Drop.

3.1 Dataset and Preprocessing

We constructed our custom dataset by selecting images containing prominent raindrop artifacts from the Raindrop Clarity dataset [7]. Our selection process yielded 488 rainy images and 80 rain-free images (negative samples). All images were manually annotated using the classification scheme in Fig. 1, yielding four categories: *large_raindrop*, *small_raindrop*, *blur*, and *streak*. The class distributions are summarized in Table 1.

Table 1: Number of Annotated Instances per Class

Scene	Streak	Blur	Large Drop	Small Drop	Total
Day	164	154	396	1,519	2,233
Night	25	46	80	176	327

To ensure strict consistency with the morphological definitions (Fig. 1), the entire dataset was

annotated by a single expert annotator. This prioritizes internal consistency at the cost of potential individual bias compared to multi-annotator consensus—a limitation addressed in Sec. 5.5. The dataset will be made available upon request for research purposes. All images were resized to 256×256 pixels to balance computational efficiency with training feasibility.

Pixel-wise Class Distribution. The pixel-wise distribution across all five classes reveals an extreme class imbalance: raindrop artifacts (Classes 1–4) collectively occupy only 0.988% of total pixels (Large Drop: 0.258%, Small Drop: 0.311%, Blur: 0.161%, Streak: 0.258%), with the Background class accounting for 99.012%. This sparsity motivates the use of Dice Loss.

3.2 Evaluation Metrics

We evaluate performance using the macro-averaged Intersection over Union (IoU) over all five classes (4 rain types + 1 background). Phase 2 uses a restricted 4-class IoU (excluding background) to expose classification difficulty more directly; the rationale is detailed in Sec. 4.

3.3 Investigated Components

We run two phases. Phase 1 fixes the training strategy through ablation, then uses it for a broad architecture and encoder search. Phase 2 takes the best Phase 1 model and studies loss composition and class granularity in detail.

- **Phase 1 – Architectures & Encoders:** We compare four architectures (U-Net, U-Net++, DeepLabV3+, PSPNet) and five encoders (MobileNetV2, ResNet34, EfficientNet-B4, ResNeXt-50, MiT-B4).
- **Phase 2 – Training Strategies:** Using the Phase 1 best model (U-Net + MiT-B4), we analyze loss function composition and classification granularity.

All models were trained for a maximum of 150 epochs with early stopping (patience=3).

4 Results and Analysis

Phase 1 uses a 5-class macro-averaged IoU (4 rain types + background) to compare architectures broadly; the large, easily-classified background raises absolute scores but is appropriate for an architecture search setting. Phase 2 switches to a 4-class IoU (raindrop classes only) to focus on the harder classification sub-problem, since the background is trivially easy to segment given its 99.0% pixel share. All scores are means over five validation folds.

4.1 Phase 1: Identifying the Optimal Model and Training Configuration

4.1.1 Ablation of Training Strategies

Using U-Net with MiT-B4 as a starting point, we tested optimizers, learning rate schedulers, and loss functions (Table 2). AdamW outperforms SGD by 0.418 IoU points (0.613 vs. 0.195),

Cosine Annealing adds 0.016 points over no scheduling, and combining CE with Dice loss adds 0.019 points over Dice alone. We fixed AdamW, Cosine Annealing, and combined CE+Dice at learning rate 1×10^{-4} for all subsequent Phase 1 runs.

Table 2: Ablation of Key Training Strategy Components

Component	Configuration	Mean IoU (5-Class)
Optimizer	AdamW + Cosine Annealing + Combined Loss	**0.613**
	SGD + Cosine Annealing + Combined Loss	0.195
LR Scheduler	AdamW + Cosine Annealing + Combined Loss	**0.613**
	AdamW + None + Combined Loss	0.597
Loss Function	AdamW + Cosine Annealing + Combined Loss	**0.613**
	AdamW + Cosine Annealing + Dice Only	0.594

4.1.2 Comparison of Architectures and Encoders

Table 3 summarizes the IoU of all architecture–encoder combinations under the fixed training strategy. U-Net with MiT-B4 achieves the highest score (0.613), and was selected as the champion model for Phase 2.

Table 3: Macro IoU of Different Architectures and Encoders (5-Class)

Architecture	MobileNetV2	ResNet34	Eff.-B4	ResNeXt-50	MiT-B4
U-Net	.506±.024	.522±.012	.540±.016	.552±.025	**.613±.020**
U-Net++	.522±.014	.533±.015	.545±.018	.573±.012	—
DeepLabV3+	.473±.018	.521±.017	.510±.026	.536±.029	.583±.015
PSPNet	.434±.015	.480±.015	.486±.013	.508±.008	.534±.020

4.2 Phase 2: Focused Analysis of the Champion Model

4.2.1 Optimization of Combined Loss Function

We evaluated the effect of varying w_{CE} and w_{Dice} in $\mathscr{L} = w_{CE}\,\mathscr{L}_{CE} + w_{Dice}\,\mathscr{L}_{Dice}$ subject to $w_{CE} + w_{Dice} = 1$ (Table 4). The setting $w_{CE} = 0.3$, $w_{Dice} = 0.7$ (Combined_CE0.3) yields the highest 4-class IoU (0.5297), a 2.5-point improvement over Dice alone.

4.2.2 Classification Granularity and Baseline Comparison

Table 5 compares four classification settings.

Table 4: Loss Function Composition Ablation (U-Net + MiT-B4, 4-Class IoU)

Experiment	w_{CE} / w_{Dice}	Mean IoU (↑)	Mean Dice (↑)
Combined_CE0.3	0.3 / 0.7	**0.5297**	**0.6897**
Combined_CE0.1	0.1 / 0.9	0.5265	0.6870
Combined_CE0.7	0.7 / 0.3	0.5274	0.6879
Combined_CE0.5	0.5 / 0.5	0.5105	0.6727
Dice Only	0.0 / 1.0	0.5043	0.6676
CE Only	1.0 / 0.0	0.4965	0.6508

Table 5: Comparison of Classification Granularity (U-Net + MiT-B4, 4-Class IoU)

Experiment	Classification	Mean IoU (↑)	Blur IoU
E2	Single Class (All Rain)	**0.6696**	N/A
E1b (CE-Prioritized)	4-Class (CE 0.7/Dice 0.3)	0.5274	0.433
E3	4-Class (Weighted Loss)	0.4821	0.376
E4	3-Class (Excl. Small Drop)	0.5061	0.443

- **Detection vs. classification gap.** The single-class baseline (E2) achieves an IoU of 0.6696, some 14 points above the best four-class result (0.5274, E1b). The model can reliably locate raindrop-like regions; the bottleneck is assigning the correct morphological label.
- **Effect of class prioritization.** E1b ($w_{CE} = 0.7$) achieves the best macro IoU among multi-class settings despite Table 4 suggesting $w_{CE} = 0.3$ is numerically superior in isolation. The stronger CE weighting improves Blur IoU in particular, which is the dominant source of error across all multi-class configurations.

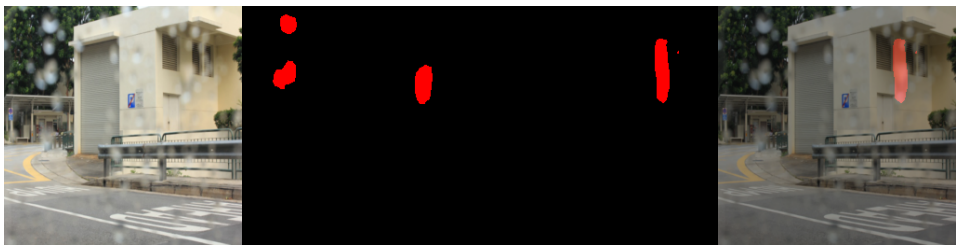
4.2.3 Generalization Test on External Dataset

We conducted a transfer test using 10 images from the RainDropRemoval dataset [8] to assess generalization to a blur-dominant distribution. This test set was constrained to 10 images because large, focused raindrops—the primary class of interest in our annotation scheme—are rare in standard synthetic open datasets, which predominantly depict rain streaks; we could not construct a larger balanced set without introducing a different distribution shift. Results are shown in Table 6.

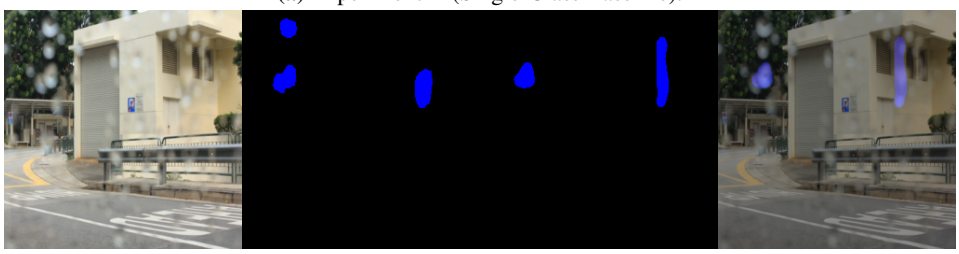
Multi-class models (E4, E1b) trail the single-class baseline in average IoU, as expected: even a correctly localized region is penalized when assigned the wrong morphological sub-label. The per-class numbers tell a different story: E4 and E1b both exceed E2 on Blur IoU (0.334 and 0.324 vs. 0.302). Multi-class training appears to encourage the encoder to learn morphology-discriminative features, and those features transfer better to the diffuse, ambiguous artifacts that dominate real-world footage.

Table 6: Generalization Test on External RainDropRemoval Dataset (Blur-Dominant)

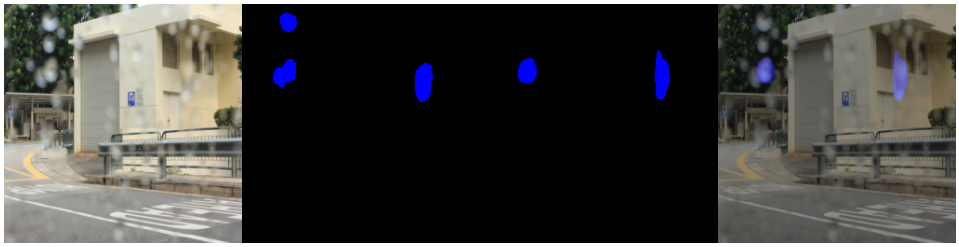
Experiment	Classification	Avg IoU (↑)	Blur IoU (↑)
E2_SingleClass_Baseline	Single Class (All Rain)	0.3015 ± 0.0452	0.3015
E4_3Class_NoSmall	3-Class (Excl. Small Drop)	**0.1114** ± 0.0173	**0.334**
E1b_MultiClass_CE07	4-Class Multi (CE 0.7/Dice 0.3)	0.0809 ± 0.0053	0.324
E1_MultiClass_Best	4-Class Multi (CE 0.3/Dice 0.7)	0.0791 ± 0.0045	0.316
E3_MultiClass_Weighted	4-Class Multi (Weighted Loss)	0.0744 ± 0.0120	0.298



(a) Experiment A (Single-Class Baseline).



(b) Experiment C (4-Class, CE 0.3/Dice 0.7).



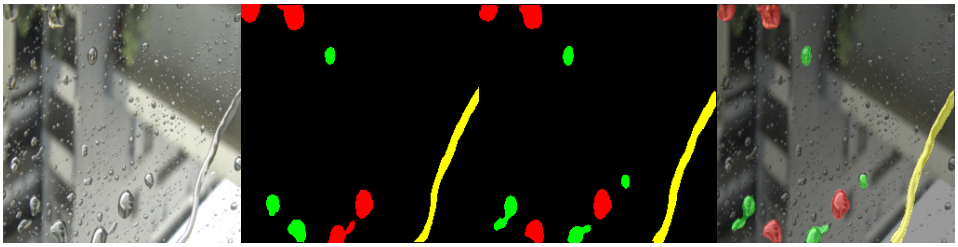
(c) Experiment D (4-Class, CE 0.7/Dice 0.3).

Fig. 3. Qualitative comparison of loss configurations. The CE-heavier setting (D) produces cleaner mask boundaries than the Dice-heavier setting (C).

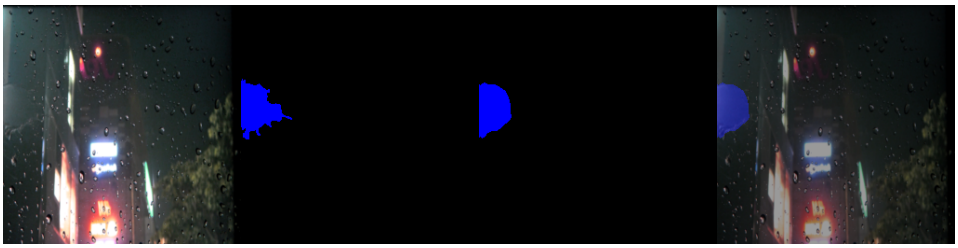
5 Discussion

5.1 Efficacy of Vision Transformers and Global Context

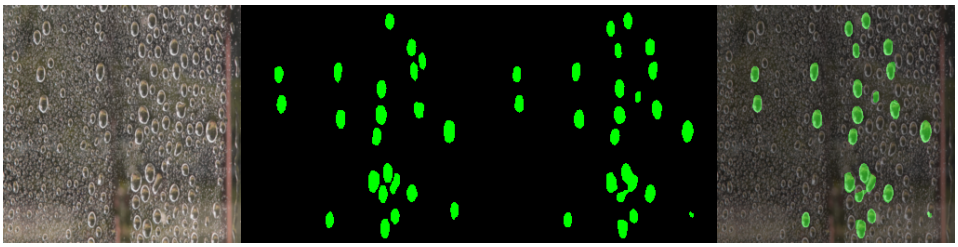
Replacing any CNN backbone with MiT-B4 raises IoU across all four architectures (Table 3); the gain is largest for U-Net (+6.1 points over ResNeXt-50). Distinguishing diffuse *Blur* from fo-



(a) Scenario 1: Clear distinction between Large Drop, Small Drop and Streak.



(b) Scenario 2: Successful segmentation of the challenging Blur class.



(c) Scenario 3: Mixed scene with a high density of small raindrops.

Fig. 4. Qualitative evaluation of the champion model (U-Net + MiT-B4).

cused *Drops* requires comparing regions that can be tens of pixels apart. MiT-B4's self-attention integrates such non-local evidence at each hierarchical stage, whereas stacked convolutions, although theoretically wide in spatial extent, concentrate their effective receptive field in a Gaussian-like pattern around the center and underweight peripheral regions [15]. A similar dependence on non-local context has been reported in medical image segmentation [13] and surface anomaly detection [14].

5.2 The Classification Bottleneck and the Blur Challenge

The gap between single-class detection (IoU 0.6696) and four-class classification (IoU 0.5274) is 14.2 points—the task's primary bottleneck is morphological labeling, not artifact localization. The *Blur* class is the main source of this difficulty. Unlike large drops, whose size threshold (≥ 50 pixels) is quantitatively defined, blur is characterized by subjective criteria such as "contour is very blurry"

or "optical details are lost." Such definitions lack a pixel-level decision boundary, making the class inherently harder for a discriminative model to learn. Even with a consistent single annotator, the annotation boundary for *Blur* is uncertain in the feature space, and Blur IoU remains the lowest among all four rain types across every experimental setting.

5.3 Loss Function Synergy and IoU-Quality Misalignment

The loss ablation in Table 4 shows that CE-only ($w_{CE} = 1.0$) scores 3.3 points below Dice-only, confirming that pixel-wise cross-entropy is overwhelmed by the 99:1 background-to-foreground ratio. Dice-only, however, produces spatially fragmented masks with salt-and-pepper noise that is invisible to the scalar IoU but detrimental for downstream deraining. This disconnect arises from a structural limitation of the standard IoU:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|}$$

The metric rewards pixel-level overlap but carries no penalty for the number of connected components in P or for topological inconsistency between P and G. We term this discrepancy *IoU-Quality Misalignment*. As Fig. 3 shows, the CE-heavier setting E1b ($w_{CE} = 0.7$) yields visually coherent boundaries at a cost of 0.2 IoU points relative to Combined_CE0.3. For a pre-processing application where mask topology affects the quality of subsequent inpainting, E1b's output is preferable despite its lower scalar score.

5.4 Enhanced Generalization for Real-World Ambiguity

On the external blur-dominant test set (Table 6), multi-class models (E4, E1b) trail the single-class baseline in average IoU, since correctly localized regions still incur a penalty for sub-label errors. On Blur IoU, however, E4 and E1b score 0.334 and 0.324 against 0.302 for the single-class model. The multi-class objective forces the encoder to distinguish morphological types, and this specialization carries over to the most ambiguous artifact class when the model is applied to an unseen distribution.

5.5 Efficiency and Computational Cost Analysis

Table 7 and Fig. 5 compare efficiency across encoders. MiT-B4 has the highest parameter count (64.12M) and FLOPs (28.30G), and runs at 86.25 FPS—lower than ResNet34 (588.61 FPS) but well above the 30 FPS real-time threshold. The IoU gain over the next-best encoder, ResNeXt-50 (0.552), is 6.1 points; for an application where missing a large raindrop has downstream consequences, this accuracy difference outweighs the throughput cost.

5.6 Limitations

The external generalization test is limited to 10 images. Large, focused raindrops—the class driving our annotation scheme—are rare in standard open datasets, which predominantly contain rain

Table 7: Comparison of Efficiency and Performance (RTX 4090)

Encoder	Params (M)	FLOPs (G)	FPS	Latency (ms)	mIoU
MobileNetV2	**6.63**	**6.86**	531.79	1.88	0.506
ResNet34	24.44	15.79	**588.61**	**1.70**	0.522
EfficientNet-B4	20.23	9.33	181.68	5.50	0.521
ResNeXt-50	31.99	21.90	465.64	2.15	0.552
MiT-B4 (Ours)	64.12	28.30	86.25	11.59	**0.613**

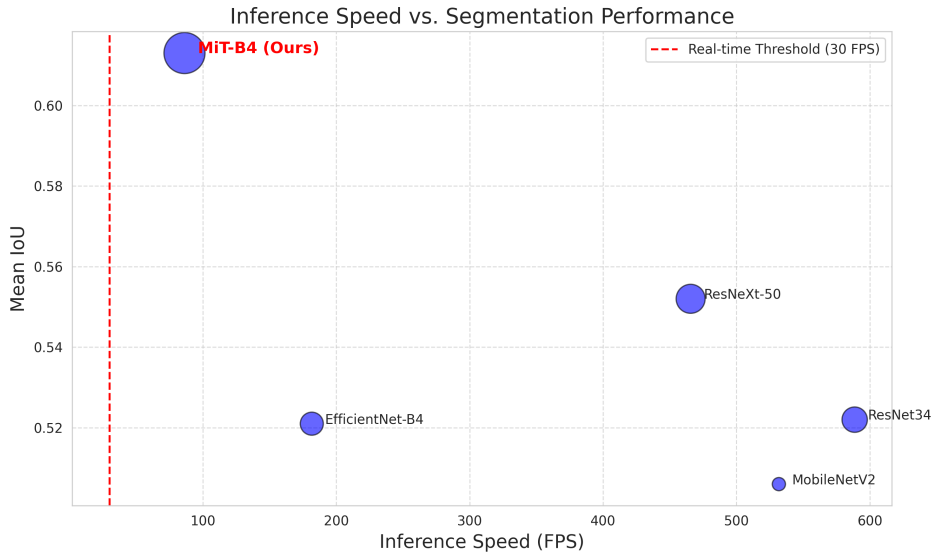


Fig. 5. Inference speed (FPS) vs. mean IoU. MiT-B4 (top right) operates well above the 30 FPS real-time threshold (red dashed line) while achieving the highest segmentation accuracy.

streaks; we could not construct a larger balanced test set without mixing distributions. Inference at 256×256 pixels may miss sub-pixel droplets in 4K footage, and future work should evaluate multi-scale inference strategies. Finally, single-annotator labeling ensures internal consistency but may introduce individual bias; multi-annotator consensus, planned for a subsequent dataset release, would provide a more principled ground truth especially for the subjective *Blur* class.

6 Conclusion

We introduced multi-class semantic segmentation of raindrops—four morphological classes motivated by the failure modes identified in our prior deraining work [9]—as a pre-processing step

for outdoor vision pipelines. Benchmarking across architectures, encoders, and training strategies identified U-Net with MiT-B4, trained with a combined CE+Dice loss under AdamW and Cosine Annealing, as the best configuration (5-class IoU: 0.613; 86.25 FPS). Two findings carry the most consequence for future work. First, the 14-point gap between single-class detection and four-class classification isolates morphological labeling—particularly of the ill-defined *Blur* class—as the task's primary bottleneck. Second, the IoU-Quality Misalignment identified in Sec. 5.3 motivates the development of topology-aware evaluation metrics for mask quality in restoration-oriented segmentation. Planned extensions include multi-annotator re-annotation to reduce individual label bias, multi-scale inference for high-resolution imagery, and temporal consistency constraints for video deployment.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[2] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolnes, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 10347–10357.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[5] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021.

[7] Y. Jin, X. Li, J. Wang, Y. Zhang, and M. Zhang, "Raindrop clarity: A dual-focused dataset for day and night raindrop removal," in *European Conference on Computer Vision (ECCV)*, pp. 1–17, 2024.

[8] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive Generative Adversarial Network for Raindrop Removal from A Single Image," *arXiv preprint arXiv:1711.10098*, 2018.

[9] K. Ichihara, R. Ishibashi, and L. Meng, "SAAFNet: Self-Adaptive Attention Fusion Network for Single Image Deraining," in *2025 International Conference on Advanced Mechatronic Systems (ICAMechS)*, 2025, pp. 120–125.

[10] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.

[11] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your ViT but faster," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[12] S. Meng et al., "AdaViT: Adaptive vision transformers for efficient image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 12309–12318.

[13] H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2022, pp. 205–218.

[14] V. Zavrtanik, M. Kruse, and D. Skocaj, "DRAEM – A discriminatively trained reconstruction embedding for surface anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 8330–8339.

[15] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2016, pp. 4898–4906.