

Edge-aware distilled segmentation with pseudo-label refinement for autonomous driving perception

Novelio Putra Indarto, Oskar Natan, Andi Dharmawan

Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia

Article Info

Article history:

Received Jul 1, 2025

Revised Oct 24, 2025

Accepted Nov 11, 2025

Keywords:

DenseCRF

Edge computing

EfficientNet

Entropy masking

Knowledge distillation

Pseudo-label refinement

Semantic segmentation

ABSTRACT

Achieving precise semantic segmentation is essential for enabling real-time perception in autonomous systems, yet leading approaches typically require substantial annotated data and powerful hardware, restricting their use on devices with limited resources. This work introduces an efficient segmentation framework that integrates pseudo-label refinement, knowledge distillation, and entropy-based confidence filtering to train compact student networks suitable for edge deployment. High-quality pseudo-labels are first produced by a robust teacher network, then further improved using a dense conditional random field to boost spatial consistency. An entropy-based selection mechanism removes unreliable predictions, ensuring that only the most trustworthy labels guide the student model's training. The use of knowledge distillation effectively transfers detailed semantic understanding from the teacher to the student, enhancing accuracy without added computational overhead. Experimental results with multiple EfficientNet backbones reveal that this pipeline improves segmentation accuracy and output clarity, while also supporting real-time or near real-time inference on CPUs with limited processing power. Extensive ablation and qualitative studies further confirm the method's robustness and flexibility for real-world edge applications.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Oskar Natan

Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences,
Universitas Gadjah Mada

Sleman Regency, Special Region of Yogyakarta, Indonesia

Email: oskarnatan@ugm.ac.id

1. INTRODUCTION

The widespread adoption of autonomous vehicles depends on robust, real-time perception systems such as semantic segmentation, which are essential for safe navigation in complex environments [1]–[5]. While state-of-the-art deep learning models like DeepLabV3+ [6], HRNet [7], and transformer-based architecture like SegFormer [8] achieve remarkable accuracy, they often require significant computational resources, making them difficult to deploy on embedded devices or consumer-grade vehicles with limited hardware capabilities [9]–[11]. To address these constraints, researchers have developed various strategies for model compression and acceleration, including network pruning [12], parameter quantization [13], and the creation of lightweight architectures like Enet [14], MobileNet [15], and EfficientNet [16]. However, lightweight models alone often struggle to maintain the detailed accuracy required for real-world, urban environments. Knowledge distillation (KD) has emerged as a powerful approach for bridging the gap between model efficiency and performance. By

transferring knowledge from a large, accurate teacher model to a smaller student, KD enables compact models to approach the accuracy of their larger counterparts with much lower resource requirements [17]. For dense prediction tasks like semantic segmentation, KD has been extended through feature imitation, attention transfer, and boundary-aware distillation [18]. Despite these advances, most KD methods still rely on the availability of annotated data, which remains a significant bottleneck due to the high cost and labor required for pixel-level labeling, especially in rapidly evolving urban scenes [18], [19]. This challenge has motivated increased interest in self-supervised, semi-supervised, and teacher-student learning frameworks that can leverage large amounts of unlabeled data [20].

A key technique in this context is pseudo-labeling, where provisional labels are assigned to unlabeled samples based on the model's confident predictions and then used to further train the network [21]. However, naïve pseudo-labeling can introduce noise and confirmation bias if unreliable predictions are included. To mitigate this, recent works use confidence thresholding, uncertainty estimation, and ensemble methods to filter ambiguous predictions [22], [23]. Spatial refinement methods like dense conditional random fields (DenseCRF) further enhance pseudo-label quality by enforcing consistency among neighboring pixels [24]. Additionally, entropy-based masking strategies allow models to dynamically select reliable regions for supervision, reducing the propagation of errors and improving overall training stability [25], [26]. Although recent advances in semi-supervised and self-supervised learning have made significant strides in reducing reliance on large labeled datasets for semantic segmentation, the majority of existing methods still require some level of manual annotation, whether for pretraining, validation, or hyperparameter tuning [27]. Moreover, ongoing research predominantly navigates controlled experimental environments where benchmarks may not effectively reflect the real-world complexities such as variable lighting, occlusions, or limited computational capacities. As a consequence, the real-world adoption of these approaches is often limited, especially for applications requiring efficient, robust, and easily deployable segmentation on hardware with strict resource constraints. In fact, only a small number of studies have specifically addressed the practical challenges of deploying segmentation models for real-time inference on edge devices with limited memory and processing power, such as those used in embedded robotics and autonomous vehicles that highlighting an ongoing need for streamlined solutions that do not depend on extensive annotation or cumbersome training procedures [28], [29].

In this work, we propose a unified, edge-aware, and fully self-supervised semantic segmentation pipeline tailored for urban autonomous driving, which eliminates the need for manual annotation at any stage. Our approach combines a transformer-based teacher for pseudo-label generation, DenseCRF for boundary refinement, entropy-based confidence masking, a lightweight EfficientNet-U-Net student, and a streamlined logit-based KD loss. We validate our framework using a real-world dataset from Universitas Gadjah Mada (UGM), Indonesia, and demonstrate effective real-time deployment on simulated embedded hardware. Our contributions include: i) a fully self-supervised pipeline for urban segmentation without manual labels, ii) a thorough analysis of pseudo-label refinement, knowledge distillation, and confidence thresholding, and iii) deployment and benchmarking on resource limited CPU or simulated edge-device conditions. Finally, we discuss the limitations and future directions for annotation-free segmentation, particularly regarding evaluation without ground truth in practical scenarios.

2. METHOD

Our proposed method addresses semantic segmentation for autonomous driving in resource-constrained urban environments, entirely without ground truth annotation. The pipeline consists of three main components: a teacher model for pseudo-label generation, a refinement step using DenseCRF to enhance label quality, and a student model trained with knowledge distillation and entropy-based confidence masking. The overall architecture is illustrated in Figure 1.

2.1. Teacher model and pseudo-Label generation

We chose SegFormer-B5 [8] as our teacher model due to its exceptional segmentation performance on urban-scene benchmarks such as Cityscapes [30]. SegFormer is the one of SOTA models that integrates Transformer-based attention mechanisms, effectively capturing global contextual information, which is critical for accurately segmenting crowded urban environments. The pre-trained teacher model generates predictions for the unlabeled images in our dataset. Specifically, we employ the teacher model as a "pseudo annotator," generating preliminary segmentation masks (pseudo-labels). The teacher outputs softmax probability maps

$P \in [0, 1]^{H \times W \times C}$, where C denotes the number of classes. The hard pseudo-labels Y at pixel (i, j) is determined by selecting the class with maximum probability.

$$Y_{i,j} = \arg \max_c P_{i,j,c} \quad (1)$$

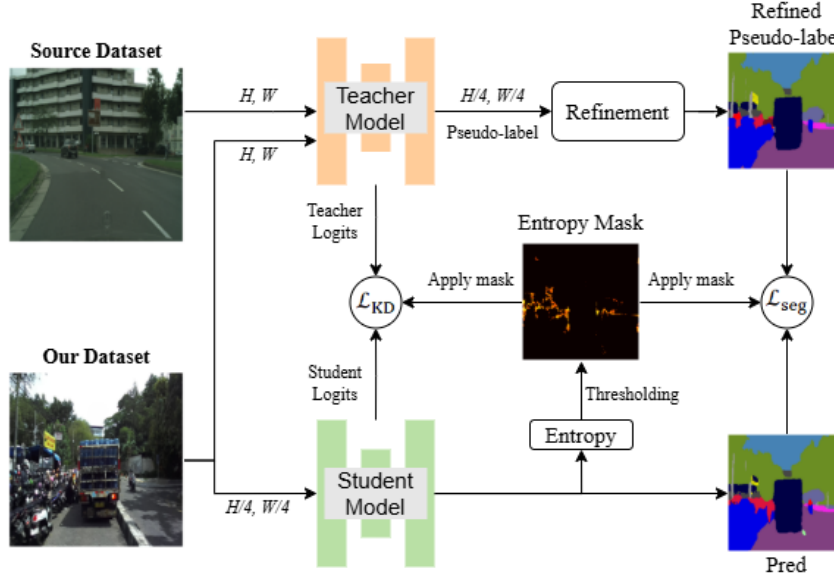


Figure 1. Complete pipeline of our method

2.2. Pseudo-label refinement with DenseCRF

While the teacher's pseudo-labels Y provide strong supervisory signals, they may contain spatial noise and misaligned boundaries. To address this, we apply DenseCRF [24] to each pseudo-label mask. DenseCRF models both unary potentials (from the teacher's predictions) and pairwise potentials that encourage spatial and appearance consistency among neighboring pixels. The energy function for the CRF is defined as (2):

$$E(\mathbf{Y}, I) = \sum_i \psi_u(Y_i | I) + \sum_{i < j} \psi_v(Y_i, Y_j | I) \quad (2)$$

where Y be the vector formed by the Y_u and Y_v pseudo-labels, where Y_u represents the unary potentials and Y_v represents the pairwise potentials. This post-processing step yields refined pseudo-labels \tilde{Y} with improved boundary accuracy and spatial coherence serve as supervision for training the student model.

2.3. Entropy-based confidence masking

Even after refinement, pseudo-labels may contain regions with high uncertainty, typically occurring at class boundaries, object occlusions, or in visually ambiguous areas. Training directly on these uncertain labels may cause the student model to learn erroneous mappings, ultimately degrading generalization and segmentation accuracy [31]. To mitigate this, we employ entropy-based confidence masking, a widely recognized approach in semi-supervised and pseudo-label learning [32]. The entropy of the teacher's predicted probability vector for pixel (i, j) is defined as:

$$H(i, j) = - \sum_c P_{i,j,c} \log P_{i,j,c} \quad (3)$$

Entropy, as a measure of uncertainty, reaches its maximum when the model is maximally uncertain and is minimized when the prediction is highly confident. We then apply a threshold τ to filter out pixels with high entropy, retaining only those with low uncertainty. The confidence mask M is defined as:

$$M(i, j) = \begin{cases} 1 & \text{if } H(i, j) < \tau \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where τ is a hyperparameter that can be tuned based on the dataset and model performance. Only pixels with $M(i, j) = 1$ are considered in loss computation, ensuring that the student model learns only from high-confidence pseudo-labels.

2.4. Student model training with knowledge distillation

The student model architecture is based on EfficientNet [16] as encoder, selected for its strong performance-to-efficiency ratio in edge and embedded applications. To enable spatially accurate segmentation, we pair this encoder with a U-Net-style decoder [33], which uses skip connections to preserve fine-grained features across multiple resolutions. The student is trained using a composite loss function that integrates two complementary objectives: a segmentation loss to enforce agreement with high-confidence, refined pseudo-labels, and a knowledge distillation loss to transfer richer semantic structure from the teacher. The total training loss is formulated as:

$$L = L_{seg} + \lambda L_{kd} \quad (5)$$

where L_{seg} is the cross-entropy loss computed only over pixels where $M(i, j) = 1$, λ is a hyperparameter balancing the two losses, and L_{kd} is the knowledge distillation loss.

2.4.1. Segmentation loss

The segmentation loss L_{seg} is defined as the sum of pixel-wise cross-entropy and dice loss between the student's predicted probability map Q and the refined pseudo-labels \tilde{Y} , weighted by the confidence mask M :

$$L_{seg} = L_{ce} + L_{dice} \quad (6)$$

$$L_{ce} = - \sum_{i,j} M(i, j) \left(\tilde{Y}(i, j) \log Q(i, j) + (1 - \tilde{Y}(i, j)) \log(1 - Q(i, j)) \right) \quad (7)$$

where $S(i, j, c)$ is the student's predicted probability for class c at pixel (i, j) .

$$L_{dice} = 1 - \frac{2 \sum_{i,j} M(i, j) \tilde{Y}(i, j) Q(i, j)}{\sum_{i,j} M(i, j) \tilde{Y}(i, j) + \sum_{i,j} M(i, j) Q(i, j)} \quad (8)$$

The Dice loss is particularly effective for handling class imbalance and improving the accuracy of object boundaries, which are critical for safety in autonomous driving scenarios.

2.4.2. Knowledge distillation loss

To further enhance the student's learning, we apply knowledge distillation (KD) in the form of Kullback-Leibler (KL) divergence between the temperature-scaled softmax outputs of the teacher and student [17]:

$$L_{kd} = \frac{1}{N} \sum_{i,j} M_{i,j} \sum_c P_{i,j,c} \log \left(\frac{P_{i,j,c}}{Q_{i,j,c}} \right) \quad (9)$$

where $P_{i,j,c}$ is the teacher's softmax output, and $Q_{i,j,c}$ is the student's softmax output. We also apply a temperature scaling factor T to soften the logits before computing the KL divergence. This loss encourages the student to mimic the teacher's distribution over classes, capturing richer semantic information beyond simple class labels.

3. RESULTS AND DISCUSSION

We evaluated the model's performance and reliability from several perspectives, including stability of the teacher model, effectiveness of loss function, qualitative outcomes, comparative analysis with baseline methods, computational efficiency, and real-world applicability through offline and online testing. To better understand the impact of each component, we conducted a qualitative analysis of the teacher model's pseudo-labels before and after refinement, as well as an ablation study on the student model's performance with and without knowledge distillation and entropy-based confidence masking across various EfficientNet backbones (B0, B3, B5, B7).

3.1. Evaluation of pseudo-label refinement

To thoroughly assess the impact of pseudo-label refinement, we conduct a series of qualitative analyses comparing the teacher's initial pseudo-labels and the results after DenseCRF post-processing. As shown in Figure 2. In this figure, each row corresponds to a different scene, with columns representing Figure 2(a) the original input image, Figure 2(b) the teacher's raw pseudo-labels, Figure 2(c) the pseudo-labels after DenseCRF refinement, and Figure 2(d) the entropy map of the teacher's softmax predictions. The yellow boxes highlight key regions of interest for closer comparison. The entropy maps clearly reveal areas of high predictive uncertainty, indicated by brighter colors. These regions tend to cluster along object boundaries, in zones with heavy occlusion, or within visually complex backgrounds—precisely the spots where the teacher model is least certain about class assignments. When we examine the corresponding pseudo-labels before and after refinement, it becomes evident that DenseCRF has its most pronounced impact in these high-entropy areas. Specifically, ambiguous or isolated pixels near boundaries are frequently “smoothed out,” with the refinement step encouraging these pixels to adopt the most prevalent local class. As a result, the refined pseudo-labels display cleaner, more continuous object shapes and boundaries, improving their suitability as supervision for student training. Despite these improvements, some challenging high-entropy regions persist, as seen in the residual label noise or ambiguous segmentation outcomes even after refinement—particularly in areas where object boundaries are especially complex or where the input image is highly cluttered. These persistent challenges underscore the limitations of spatial refinement alone and suggest that further advances in label denoising or uncertainty modeling may be necessary for even greater segmentation fidelity.

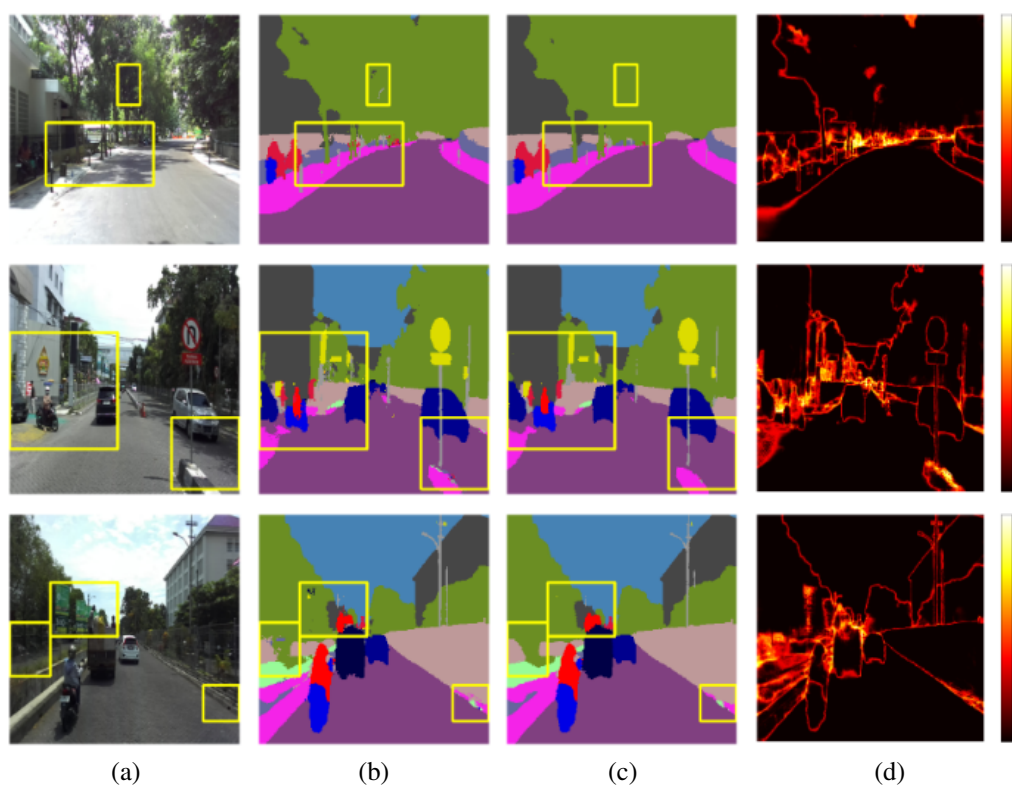


Figure 2. Qualitative analysis of pseudo-label refinement (a) input image, (b) teacher's pseudo-labels, (c) refined pseudo-labels after DenseCRF, and (d) entropy map of teacher's predictions

Comparing student performance when trained on pseudo-labels before and after applying DenseCRF. As summarized in Table 1, the use of refined pseudo-labels consistently results in improved mIoU and F1 Score across all models. For instance, EfficientNet-B0's mIoU improves from 81.89% to 82.57%, and its F1 Score from 89.97% to 90.37% with refinement, while the largest evaluated model EfficientNet-B7 achieves the best

overall metrics with 83.21% mIoU and 90.76% F1 after refinement. These gains highlight how the refinement process enhances not just the visual clarity of the labels, but also leads to more effective model training and generalization. Nevertheless, refinement is not universally beneficial in every scenario. As illustrated by the examples in Figure 3, with Figure 3(a) shows the original input image, Figure 3(b) shows the teacher's raw pseudo-labels, and Figure 3(c) shows the pseudo-labels after DenseCRF refinement, there are cases where DenseCRF refinement fails to preserve small or uncommon objects. In the highlighted regions, rare classes such as traffic poles or signs are often subsumed by the majority class, particularly when these objects occupy a small number of pixels or are situated in visually complex or occluded areas. This outcome stems from the local context-driven nature of DenseCRF, which can inadvertently erase minority details in favor of smoother segmentations.

Table 1. Enhanced performance of student model with and without refinement

Backbone	Without Refinement		With Refinement	
	mIoU	F1 Score	mIoU	F1 Score
EfficientNet-B0	81.89	89.97	82.57	90.37
EfficientNet-B3	81.52	89.74	82.75	90.49
EfficientNet-B5	82.24	90.18	82.79	90.50
EfficientNet-B7	82.70	90.45	83.21	90.76

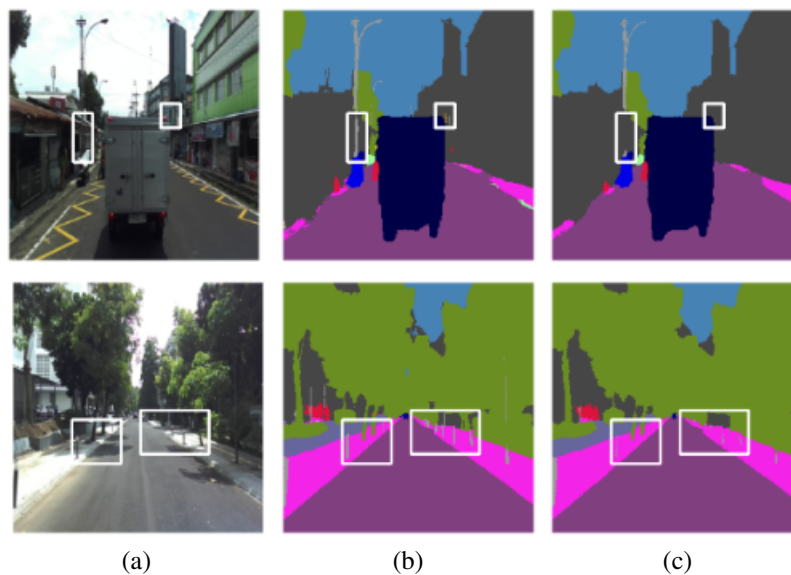


Figure 3. Failure cases where refinement absorbed rare or small objects (a) input image, (b) teacher's pseudo-labels, and (c) refined pseudo-labels after DenseCRF

3.2. Model performance evaluations

To rigorously assess our segmentation pipeline, we systematically evaluated the impact of knowledge distillation (KD) and entropy-based masking on segmentation performance using refined pseudo-labels, focusing on mean Intersection over Union (mIoU) and F1 score as primary metrics. As summarized in Table 2, incorporating consistently improved both mIoU and F1 scores across all backbone variants. For EfficientNet-B0, the smallest and most resource-efficient model, there is a clear synergy when combining knowledge distillation with entropy-based filtering shows that the highest mIoU (82.90%) and F1 Score (90.58%) are achieved when the student is guided by the teacher and simultaneously shielded from low-confidence pseudo-labels at an entropy threshold of 0.45. This suggests that, for lightweight models, restricting learning to confident predictions is essential for preventing overfitting to noise, while knowledge distillation provides the semantic richness necessary to approach teacher-level accuracy even in edge-limited settings.

The narrative is more complex with EfficientNet-B3 and B5. Notably, knowledge distillation alone,

without any entropy masking, produces little to no improvement over the baseline, indicating that these mid-sized models cannot reliably extract value from teacher supervision in the presence of label noise. However, when knowledge distillation is paired with increasing entropy thresholds, both B3 and B5 display a marked increase in performance, peaking at 82.95% mIoU and 90.61% F1 for B3 at ET 0.45, and 83.00% mIoU and 90.64% F1 for B5 at ET 0.35. These results highlight that, for mid-capacity architectures, the combination of robust, high-confidence pseudo-labels and teacher-driven semantic guidance is key to achieving optimal results. This also underscores the sensitivity of these models to the quality of supervision that too much noisy or ambiguous data can obscure the benefits of knowledge transfer.

Table 2. Model performance with and without KD and ET

Backbone	ET	Without KD		With KD	
		mIoU	F1 Score	mIoU	F1 Score
EfficientNet-B0	-	82,57	90,37	82,86	90,55
	0.05	81,52	89,74	82,72	90,47
	0.15	81,92	89,98	82,71	90,45
	0.25	82,27	90,19	82,68	90,44
	0.35	82,02	90,04	82,66	90,43
	0.45	82,16	90,13	82,90	90,58
EfficientNet-B3	-	82,75	90,49	82,37	90,25
	0.05	82,37	90,25	82,75	90,49
	0.15	82,17	90,14	82,23	90,17
	0.25	82,56	90,37	82,63	90,41
	0.35	82,65	90,42	82,83	90,53
	0.45	82,81	90,52	82,95	90,61
EfficientNet-B5	-	82,79	90,50	82,59	90,39
	0.05	82,23	90,17	82,90	90,57
	0.15	82,58	90,38	82,95	90,60
	0.25	82,06	90,06	82,95	90,61
	0.35	82,49	90,32	83,00	90,64
	0.45	82,55	90,36	82,91	90,58
EfficientNet-B7	-	83,21	90,76	83,88	91,16
	0.05	82,66	90,43	83,61	91,00
	0.15	82,99	90,36	83,71	91,06
	0.25	83,47	90,91	83,78	91,10
	0.35	83,10	90,69	83,66	91,03
	0.45	83,27	90,79	83,53	90,95

The impact of entropy-based confidence masking, however, depended strongly on model capacity. For smaller backbones such as B0 and B3, entropy masking further improved segmentation accuracy by filtering out high-uncertainty pseudo-labels, thus reducing the risk of overfitting to noise, a trend that was most evident at moderate threshold values. This suggests that lightweight models, being more vulnerable to label noise, benefit from selective training on reliable supervision. In contrast, a different pattern emerged with the high-capacity EfficientNet-B5. Omitting entropy masking and utilizing all pseudo-labels, including those with greater uncertainty, led to the highest segmentation metrics. This finding indicates that larger models can tolerate or even benefit from exposure to ambiguous regions, likely due to their superior ability to extract informative signals and generalize from a more diverse set of training examples.

For EfficientNet-B7, the largest backbone evaluated, a different trend emerges. The model achieves its best performance (83.88% mIoU and 91.16% F1) with knowledge distillation alone and no entropy thresholding, and sustains high metrics even as entropy thresholds are varied. This resilience suggests that high-capacity models are equipped to handle the inherent noise in pseudo-labels, capitalizing on their greater representational power and capacity to learn from a diverse and even ambiguous set of training signals. For B7, additional filtering may actually reduce the beneficial exposure to challenging, informative regions that contribute to fine-grained segmentation accuracy.

Taken together, these results reveal an important interplay between model capacity, supervision quality, and deployment considerations. While knowledge distillation is a universally valuable tool for narrowing the performance gap between student and teacher models, but its effectiveness is highly dependent on how supervision is delivered and the underlying capacity of the model. Applying entropy-based masking can significantly enhance performance for smaller models by ensuring they learn from high-confidence signals, but

may not be necessary or even beneficial for larger architectures that can handle more noise. This suggests that the decision to apply entropy-based masking should be guided by the backbone's representational power and the intended resource constraints of the deployment environment. Overall, our experiments demonstrate that the pipeline is both flexible and effective, offering strong segmentation accuracy and efficient inference for edge-aware applications.

3.3. Edge-aware evaluation

To evaluate the suitability of our segmentation pipeline for real-world edge deployment, we conducted an extensive resource-awareness study by simulating inference on a CPU. We use AMD Ryzen Threadripper PRO 7965WX CPU with limited 1-4 cores configurations. For each setting, we measured the student model's computational requirements and inference efficiency using the following metrics: average inference time per image, frames per second (FPS), RAM usage, and GFLOPS. Results are summarized in Table 3.

Table 3. Edge-Aware Evaluation of Student Models with Varying Cores

Backbone	Metric	1 Core	2 Cores	3 Cores	4 Cores
EfficientNet-B0 Size (MB): 68.95 Parameters (M): 5.80	Inferen Time (ms)	63.93 ± 2.03	38.02 ± 2.36	28.36 ± 1.72	23.83 ± 1.05
	FPS	15.66 ± 0.48	26.39 ± 1.54	35.37 ± 1.94	42.04 ± 1.81
	RAM (MB)	1015.11 ± 41.79	1003.92 ± 37.37	1008.54 ± 33.75	1014.72 ± 35.47
	GFLOPS	98.45 ± 3.01	165.96 ± 9.68	222.43 ± 12.20	264.36 ± 11.38
EfficientNet-B3 Size (MB): 146.97 Parameters (M): 12.48	Inferen Time (ms)	89.86 ± 2.05	55.03 ± 1.93	42.21 ± 1.50	34.76 ± 1.22
	FPS	11.13 ± 0.26	18.19 ± 0.65	23.72 ± 0.85	28.80 ± 1.00
	RAM (MB)	1124.24 ± 40.36	1106.92 ± 31.21	1116.28 ± 27.26	1125.34 ± 47.58
	GFLOPS	89.14 ± 2.07	145.66 ± 5.17	189.91 ± 6.78	230.60 ± 8.01
EfficientNet-B5 Size (MB): 350.67 Parameters (M): 30.00	Inferen Time (ms)	132.28 ± 2.03	77.44 ± 0.41	60.34 ± 1.12	49.78 ± 0.44
	FPS	7.56 ± 0.12	12.91 ± 0.07	16.58 ± 0.30	20.09 ± 0.18
	RAM (MB)	1280.45 ± 51.11	1365.62 ± 64.42	1340.14 ± 57.63	1369.87 ± 41.35
	GFLOPS	90.24 ± 1.39	154.11 ± 0.83	197.84 ± 3.56	239.75 ± 2.12
EfficientNet-B7 Size (MB): 757.59 Parameters (M): 65.15	Inferen Time (ms)	213.47 ± 0.95	119.06 ± 0.41	90.01 ± 0.51	74.56 ± 0.47
	FPS	4.68 ± 0.02	8.40 ± 0.03	11.11 ± 0.06	13.41 ± 0.08
	RAM (MB)	1854.54 ± 149.75	1751.95 ± 85.23	1896.44 ± 137.14	1840.15 ± 133.82
	GFLOPS	91.99 ± 0.41	164.93 ± 0.56	218.16 ± 1.22	263.39 ± 1.66

The results indicate that the EfficientNet-B0 model is enough to meet autonomous driving minimum FPS (10 FPS which shows as green cells in the table) [34], achieved an inference time of 63.93 ms and 15.66 FPS on a single core, improving to 23.83 ms and 42.04 FPS with four cores. Different from the EfficientNet-B0, the EfficientNet-B5 and B7 models, which are larger and more complex, require more computational resources, achieving 132.28 ms and 7.56 FPS on a single core, but still reaching more than 10 FPS with two and three cores. The EfficientNet-B3 model, which balances performance and efficiency, achieves 89.86 ms and 11.13 FPS on a single core, improving to 34.76 ms and 28.80 FPS with four cores.

Throughout all configurations, RAM usage remained largely stable for each model, ranging from approximately 1.0 to 1.8 GB indicating that increased CPU parallelism primarily affects computational speed rather than memory footprint. GFLOPS values and parameter counts reflect each backbone's inherent complexity, EfficientNet-B0 is the most lightweight at 5.8M parameters and 264.36 GFLOPS (4 cores), while B7 is the largest at 65.15M parameters and 263.39 GFLOPS (4 cores). Model sizes on disk increase accordingly, from 68.95 MB (B0) up to 757.59 MB (B7), reinforcing the trade-off between segmentation accuracy and resource demand.

Taken together with our quantitative benchmarking and qualitative analysis, these edge-aware deployment results reinforce the adaptability and robustness of our proposed pipeline. The ability to scale from lightweight, real-time models for constrained devices up to higher-capacity backbones for more demanding platforms, while maintaining strong segmentation accuracy and efficient use of computational resources, highlights the practical value of combining pseudo-label refinement, knowledge distillation, and entropy-based masking. Our comprehensive evaluation, spanning performance metrics, visual quality, and hardware efficiency, positions this approach as a flexible and effective solution for semantic segmentation in diverse real-world edge scenarios.

4. CONCLUSION

In this paper, we presented a novel edge-aware semantic segmentation pipeline that leverages pseudo-label refinement, knowledge distillation, and entropy-based confidence masking to train lightweight student models for deployment on resource-constrained devices. Our approach systematically addresses the challenges of noisy supervision and model capacity, enabling effective segmentation in real-world scenarios with limited annotations. We demonstrated that our pipeline can achieve strong segmentation performance across various EfficientNet backbones, with significant improvements in mIoU and F1 scores through the use of refined pseudo-labels and knowledge distillation. The entropy-based confidence masking further enhances model robustness by filtering out uncertain labels, particularly benefiting smaller models that are more susceptible to noise. We also conducted extensive edge-aware evaluations, showing that our pipeline can deliver real-time or near real-time inference on CPUs with limited core counts, making it suitable for deployment in autonomous driving and other edge computing applications. The results indicate that even lightweight models can achieve competitive segmentation accuracy while maintaining low computational overhead and memory requirements, thus enabling practical deployment on embedded and IoT devices.

Overall, our findings highlight the importance of harmonizing model capacity with supervision quality and deployment constraints, paving the way for future research on adaptive label refinement, advanced knowledge transfer techniques, and context-aware deployment strategies in semantic segmentation for edge computing environments. By addressing the challenges of noisy labels and resource limitations, our pipeline contributes to the development of efficient and effective segmentation solutions that can be deployed in real-world applications, particularly in autonomous driving scenarios where safety and reliability are paramount.

FUNDING INFORMATION

This work is supported by Universitas Gadjah Mada under the funding of the Final Project Recognition Program (RTA, contract number: 5286/UN1.P1/PT.01.03/2024) for the Year 2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Novelio Putra Indarto	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	
Oskar Natan		✓				✓	✓	✓	✓	✓		✓	✓	✓
Andi Dharmawan		✓			✓		✓		✓	✓		✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal Analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project Administration

Fu : Funding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, ON. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

REFERENCES




- [1] O. Natan and J. Miura, "Towards compact autonomous driving perception with balanced learning and multi-sensor fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 16249–16266, Sep. 2022, doi: 10.1109/TITS.2022.3149370.

- [2] K. Ishihara, A. Kanervisto, J. Miura, and V. Hautamäki, "Multi-task learning with attention for end-to-end autonomous driving," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2021, pp. 2896–2905, doi: 10.1109/CVPRW53098.2021.00325.
- [3] O. Natan and J. Miura, "Semantic segmentation and depth estimation with RGB and DVS sensor fusion for multi-view driving perception," in *Asian Conference on Pattern Recognition*, Nov. 2022, pp. 352–365, doi: 10.1007/978-3-031-02375-0_26.
- [4] H. K. Chiu, E. Adeli, and J. C. Niebles, "Segmenting the future," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4202–4209, Jul. 2020, doi: 10.1109/LRA.2020.2992184.
- [5] O. Natan and J. Miura, "End-to-end autonomous driving with semantic depth cloud mapping and multi-agent," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 557–571, Jan. 2023, doi: 10.1109/TIV.2022.3185303.
- [6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Lecture Notes in Computer Science*, vol. 11211, pp. 833–851, 2018, doi: 10.1007/978-3-030-01234-2_49.
- [7] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021, doi: 10.1109/TPAMI.2020.2983686.
- [8] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Álvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 15, pp. 12077–12090, 2021.
- [9] H. Bolhasani and S. J. Jassbi, "Deep learning accelerators: A case study with MAESTRO," *Journal of Big Data*, vol. 7, no. 1, p. 100, 2020, doi: 10.1186/s40537-020-00377-8.
- [10] M. Xu, Y. Gu, S. Poslad, and S. Xue, "Optimized computation combining classification and detection networks with distillation," in *International Joint Conference on Neural Networks*, 2021, pp. 1–8, doi: 10.1109/IJCNN52387.2021.9534331.
- [11] Z. Yang, "Synergies and challenges in the integration of cloud computing and deep learning: Current status, interconnectedness, and future directions," *Highlights in Science, Engineering and Technology*, vol. 97, pp. 100–105, 2024, doi: 10.54097/70266446.
- [12] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the value of network pruning," *International Conference on Learning Representations*, 2019.
- [13] B. Jacob *et al.*, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2704–2713, doi: 10.1109/CVPR.2018.00286.
- [14] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," *arXiv:1606.02147*, 2016.
- [15] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861*, 2017.
- [16] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, 2019, pp. 10691–10700.
- [17] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv:1503.02531*, 2015.
- [18] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.
- [19] X. Liu, M. Neuyen, and W. Q. Yan, "Vehicle-related scene understanding using deep learning," in *Communications in Computer and Information Science*, vol. 1180, pp. 61–73, 2020, doi: 10.1007/978-981-15-3651-9_7.
- [20] Y. Chen, W. Li, and L. Van Gool, "ROAD: Reality oriented adaptation for semantic segmentation of urban scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7892–7901, doi: 10.1109/CVPR.2018.00823.
- [21] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *International Joint Conference on Neural Networks*, 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9207304.
- [22] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," *International Conference on Learning Representations*, 2021.
- [23] L. Lu, M. Yin, L. Fu, and F. Yang, "Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation," *Biomedical Signal Processing and Control*, vol. 79, p. 104203, 2023, doi: 10.1016/j.bspc.2022.104203.
- [24] A. Arnab *et al.*, "Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 37–52, 2018, doi: 10.1109/MSP.2017.2762355.
- [25] Y. Wang, Y. Huang, Q. Wang, C. Zhao, Z. Zhang, and J. Chen, "Graph-based self-training for semi-supervised deep similarity learning," *Sensors*, vol. 23, no. 8, p. 3944, 2023, doi: 10.3390/s23083944.
- [26] T. Cheng, K. Lu, H. Wang, Y. Zhao, C. Jia, and J. Xue, "Semi-supervised medical image segmentation based on cross-pseudo-supervised guidance with high-confidence pseudo labeling," in *Medical Imaging 2025: Image Processing*, vol. 13406, SPIE, 2025, pp. 546–553, doi: 10.1117/12.3047161.
- [27] Q. Jiang *et al.*, "Weakly-supervised image semantic segmentation based on superpixel region merging," *Big Data and Cognitive Computing*, vol. 3, no. 2, p. 31, 2019, doi: 10.3390/bdcc3020031.
- [28] M. A. M. Elhassan *et al.*, "Real-time semantic segmentation for autonomous driving: A review of CNNs, Transformers, and beyond," *Journal of King Saud University – Computer and Information Sciences*, vol. 36, no. 10, p. 102226, 2024, doi: 10.1016/j.jksuci.2024.102226.
- [29] O. Natan and J. Miura, "DeepIPC: Deeply integrated perception and control for an autonomous vehicle in real environments," *IEEE Access*, vol. 12, pp. 49590–49601, Apr. 2024, doi: 10.1109/ACCESS.2024.3385122.
- [30] M. Cordts *et al.*, "The Cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp. 3213–3223, doi: 10.1109/CVPR.2016.350.
- [31] S. Matsuzaki, J. Miura, and H. Masuzawa, "Multi-source pseudo-label learning of semantic segmentation for the scene recognition of agricultural mobile robots," *Advanced Robotics*, vol. 36, no. 19, pp. 1011–1029, 2022, doi: 10.1080/01691864.2022.2109427.
- [32] M. J. Meni, R. T. White, M. L. Mayo, and K. R. Pilkievicz, "Entropy-based guidance of deep neural networks for accelerated




- convergence and improved performance,” *Information Sciences*, vol. 681, p. 121239, 2024, doi: 10.1016/j.ins.2024.121239.
- [33] B. Baheti, S. Innani, S. Gajre, and S. Talbar, “Eff-UNet: A novel architecture for semantic segmentation in unstructured environment,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 1473–1481, doi: 10.1109/CVPRW50498.2020.00187.
- [34] X. Wang, M. A. Maleki, M. W. Azhar, and P. Trancoso, “Moving forward: A review of autonomous driving software and hardware systems,” *arXiv:2411.10291*, 2024.

BIOGRAPHIES OF AUTHORS






Novelio Putra Indarto    received his B.Sc. degree in electronics and instrumentation from Universitas Gadjah Mada, Indonesia, in 2023, and is currently pursuing an M.Sc. degree in the same field at Universitas Gadjah Mada. His research focuses on advancing intelligent systems and automation, including multi-coordination drones and autonomous driving AI models. He has also conducted research on diverse topics, such as developing an electronic nose for detecting hazardous compounds, designing a driver safety system to monitor drowsiness and alcohol levels, and exploring horizon detection techniques for stabilizing unmanned aerial vehicles (UAVs). With a strong foundation in robotics and machine learning, Novelio aims to address real-world challenges through innovative automation technologies. He can be contacted at email: novelio.p.i@mail.ugm.ac.id.



Oskar Natan    received the B.A.Sc. degree in electronics engineering and the M.Eng. degree in electrical engineering from Politeknik Elektronika Negeri Surabaya, Indonesia, in 2017 and 2019, respectively, and the Ph.D. (Eng.) degree in computer science and engineering from Toyohashi University of Technology, Japan, in 2023. Since January 2020, he has been affiliated with the Department of Computer Science and Electronics, Universitas Gadjah Mada, Indonesia, first as a lecturer and currently an assistant professor. His research interests include sensor fusion, hardware acceleration, and end-to-end systems. He is a member of the IEEE ITS Society, the IEEE-RA Society, and the Indonesian Computer, Electronics, and Instrumentation Support Society (IndoCEISS). He has been serving as a reviewer for some reputable journals and conferences, including IEEE Transactions on Intelligent Vehicles, IEEE Transactions on Intelligent Transportation Systems, IEEE ICRA, and IEEE/RSJ IROS. He can be contacted at email: oskarnatan@ugm.ac.id.



Andi Dharmawan    received the B.Sc. degree in electronics and instrumentation, M.Cs. degree in computer science, and Doctor in computer science from Universitas Gadjah Mada in 2006, 2009, and 2017, respectively. His research interests include unmanned aerial vehicles (UAVs), control systems, and robotics. In addition to his role as a lecturer, Andi Dharmawan serves as the supervisor of the Gadjah Mada Flying Object Research Center and the Gadjah Mada Robotics Team, fostering innovation and development in aerospace and robotics research. He also holds the position of Secretary in the Department of Computer Science and Electronics at Universitas Gadjah Mada. He is a member of the IEEE society and the Indonesian Computer, Electronics, and Instrumentation Support Society (IndoCEISS). He can be contacted at andi_dharmawan@ugm.ac.id.